Studies on the Salient Properties of Digital Imagery that Impact on Human Target Acquisition and the Implications for Image Measures.

A thesis submitted for the degree of

DOCTOR OF PHILOSOPHY

in

The Departments of Computer Science & Psychology,
The University of Adelaide, South Australia.

by

Gary John Ewing, B.App.Sc. M.Sc.

January 11, 1999

Contents

List of Figures

List of Tables

Declaration

I declare that this thesis is a record of original work and that it contains no material which has been accepted for the award of any other degree or diploma in any University.

To the best of my knowledge and belief, this thesis contains no material previously published or written by any other person, except where due reference is given in the text of the thesis.

I consent to this thesis being made available for photocopying or loan.

Gary J. Ewing

January 1999

Acknowledgements

I wish to express thanks to my academic supervisors.

I thank Dr Michael Brooks of the Department of Computer Science, for accepting me as a PhD candidate and for our early discussions.

I thank Dr Chris Woodruff of the Defence Science & Technology Organisation (DSTO) for offering his expertise in visual psychophysical experimentation and for our many fruitful discussions. Dr Woodruff's guidance came at a time when I needed focusing, and I greatly appreciated his support.

I thank Dr Douglas Vickers of the Department of Psychology, for offering to become my principal academic supervisor, at a time when I realised my work fell largely under his academic discipline. I have relied on Dr Vickers in assessing my thesis as suitable for submission.

I thank Dr Nicholas Redding of DSTO for his encouragement and conscientiousness in spurring me on. He proof read my thesis drafts, with particular emphasis on the mathematical components, which was very helpful. Dr Redding gave considerable input into the development of my work discussed in Chapter 9.

I thank Dr Leighton Barnden, of the Department of Nuclear Medicine in the Queen Elizabeth Hospital (QEH) Adelaide, for his help and encouragement in carrying out the work in Chapter 8. I also appreciated the friendship he extended to me during my stay at the QEH.

I wish to thank others for contributions in some way to my work.

I thank Dr Garry Newsam of DSTO for his very valuable comments during the course of my early experiments from inception to completion and for the use of his smooth zooming algorithm.

I thank the DSTO management for their support in allowing me to enrol for a PhD on a half time basis and facilitating my use of DSTO resources and time. In particular, I would like to thank Dr Roger Lough, Chief Land Operations Division, for granting general and financial support. I also thank Dr Tin French and Dr Jeremy Manton for allowing me time off to write up my thesis.

I thank my family for their support in keeping me going and in particular I thank my wife Barbara, to whom this thesis is dedicated. Her encouragement and loving devoted support allowed this work to reach fruition.

Finally, I like to acknowledge, that this thesis was completed through the Grace of God.

Work Performed by Others.

All of the work discussed in this thesis, including experimental design, set up and running of experiments, computer coding and analysis has been performed solely by the author except in the following cases. Of course as has been already acknowledged, the work benefited from

discussions with my PhD supervisors.

I acknowledge the technical support of Mr Warwick Holen, who wrote the code to allow software control of the MPEG-2 play-back board used in the video experiments of Chapter 6 and who helped with video pre-processing.

I acknowledge the programming and technical support of Mr David Kettler, who developed the experimental software used for the work discussed in Chapter 9.

I acknowledge the assistance of Dr Philip Chapple, who under my suggestion, developed the Matlab code for the fractal simulation of image clutter used in Chapter 7.

Abstract

Electronically displayed images are becoming increasingly important as an interface between man and information systems. Lengthy periods of intense observation are no longer unusual. There is a growing awareness that specific demands should be made on displayed images in order to achieve an optimum match with the perceptual properties of the human visual system. These demands may vary greatly, depending on the task for which the displayed image is to be used and the ambient conditions. Optimal image specifications are clearly not the same for a home TV, a radar signal monitor or an infra-red targeting image display. There is, therefore, a growing need for means of objective measurement of image quality, where "image quality" is used in a very broad sense and is defined in the thesis, but includes any impact of image properties on human performance in relation to specified visual tasks.

The aim of this thesis is to consolidate and comment on the image measure literatures, and to find through experiment the salient properties of electronically displayed real world complex imagery that impacts on human performance. These experiments were carried out for well specified visual tasks (of real relevance), and the appropriate application of image measures to this imagery, to predict human performance, was considered.

An introduction to certain aspects of image quality measures is given, and clutter metrics are integrated into this concept. A very brief and basic introduction to the human visual system (HVS) is given, with some basic models. The literature on image measures is analysed, with a resulting classification of image measures, according to which features they were attempting to quantify.

A series of experiments were performed to evaluate the effects of image properties on human performance, using appropriate measures of performance. The concept of image similarity was explored, by objectively measuring the subjective perception of imagery of the same scene, as obtained through different sensors, and which underwent different luminance transformations. Controlled degradations were introduced, by using image compression. Both still and video compression were used to investigate both spatial and temporal aspects of HVS processing. The effects of various compression schemes on human target acquisition performance were quantified. A study was carried out to determine the "local" extent, to which the clutter around a target, affects its detectability. It was found in this case, that the excepted wisdom, of setting the local domain (support of the metric) to twice the expected target size, was incorrect. The local extent of clutter was found to be much greater, with this having implications for the application of clutter metrics. An image quality metric called the *gradient energy measure* (GEM), for quantifying the affect of filtering on Nuclear Medicine derived images, was developed and evaluated. This proved to be a reliable measure of image smoothing and noise level, which in preliminary studies agreed with human perception The final study discussed in this thesis determined the performance of human image analysts, in terms of their receiver operating characteristic, when using Synthetic Aperture Radar (SAR) derived images in the surveillance context. In particular, the effects of target contrast and background clutter on human analyst target detection performance were quantified. In the final chapter, suggestions to extend the work of this thesis are made, and in this context a system to predict human visual performance, based on input imagery, is proposed. This system intelligently uses image metrics based on the particular visual task and human expectations and human visual system performance parameters.

Publications and Abstracts

During the course of my study the following papers and reports have been published, or have been presented at learned society conferences.

- "A Comparison of JPEG and Fractal Based Image Compression on Target Acquisition by Human Observers." G.J. Ewing and C.J. Woodruff, *Optical Engineering*, Vol 35(1), pp 284-288, 1996.
- "Studies on the Effects of Image & Video Compression on Human Visual Task Performance" Gary Ewing and Chris Woodruff, *The Proceedings of the International Picture Coding symposium* (PCS'96), April 1996.
- "Image Analyst's Performance in Search Mode Target Detection with SAR Imagery" Gary J. Ewing, Nicholas J. Redding and David I. Kettler *The proceedings of the International workshop on Image Analysis and Image Fusion* (IAIF '97)(IEEE).
- "Effects of Line Source and Processing Factors on Results of Simultaneous TI-201 Emission and Tc-99m Fan beam Transmission Tomography" L.R. Barnden, G.J. Ewing and J.A. McKinnon Presented at the *World Congress in Nuclear Medicine*, Sydney Australia, October 1994.
- "An Image Measure for Use in Optimizing SPECT Images for Visual Diagnoses: the Gradient Energy Measure" Gary Ewing and Leighton Barnden Presented at the *Australian conference on Nuclear Medicine & Technology* at Brisbane in May 1995.
- "Towards the Visual Optimisation of Pre-filtering for SPECT images" Gary Ewing Queen Elizabeth Hospital report Jan 1995

  DSTO Reports:

- "Assessment of Unmanned Aerial Vehicles Video : Suitability For Compression Experiments and Broad Directions for Research" G. Ewing ITD report no. 94-01, 1994
- "The Effect of Image Compression on Target Acquisition by Human Observers." G. J. Ewing, C. W. Woodruff and G. N. Newsam, ERL report no. ERL-0815-RR, 1994.
- "Studies of the Effect of Video Compression on Ship Recognition" G. J. Ewing Report no. LSOD 97-7-CR, 1997.
- " Image Analyst's Performance in Search Mode Target Detection for Broad Area Aerial Surveillance " G. J. Ewing, N. J. Redding and D. I. Kettler. Research Report no. DSTO-RR-0110, 1997.

Publications in Preparation

- "Studies on the Effects of Video Compression on Target Recognition" G.J. Ewing and C.W. Woodruff (journal paper).
- "The Effects of Clutter on Human Target Detection Performance" G.J. Ewing, N.J. Redding and D.J. Kettler (journal paper).
- "The Effect of 'Local' Clutter on Human Target Detection Performance" G.J. Ewing, C.W. Woodruff and D. Vickers (journal paper).

Earlier Publications

- "Reduction of Some Environmental Effects that Degrade the Performance of HF Skywave Radars" Netherway, D.J.; Ewing, G.J. and Anderson S.J. *IEEE Symposium on Signal Processing And Applications*, 17 - 19 April 1989, Adelaide.
- "A Non-invasive Study of the Third Heart sound in Children by Phono and Echocardiography". G. Ewing, J. Mazumdar, N.Fazzalari, E. Goldblatt and E. Van Vollenhoven. *Third European conference on Mechanocardiography*, Sept. 21-23 1983, Berlin, G.D.R.
- "A Spectral Analysis of the Third Heart Sound in Children" G. Ewing, J. Mazumdar, N.Fazzalari, E. Goldblatt *Australian National Conference on Engineering and Physics in the Life Sciences.* Aug. 21- 24 1984, Adelaide.
- An invited paper of the same title and authors as above was published in *Acta Cardiologica* vol.39(4):241-254,1984
- "Use of the Maximum Entropy Method for Spectral Analysis of the Third Heart Sound in Children" G. Ewing, J. Mazumdar, N.Fazzalari, E. Goldblatt and E. Van Vollenhoven. *Fourth European Conference on Mechanocardiography* Sept. 11-14,1985 Budapest.
- " A Comparative Study of the Maximum Entropy Method and the Fast Fourier Transform for the Spectral Analysis of the Third Heart Sound in Children " G. Ewing, J. Mazumdar, B. Vojdani, E. Goldblatt and E. Van Vollenhoven. *Australasian Physical & Engineering Sciences in Medicine* (1986) vol. 9 No. 3

Reports I authored while at the The University Of Melbourne:

- "Electrode Dissolution and Corrosion Studies" published in a report to the Dept. of Science and Technology and the U.S. Food and Drug Administration (FDA).
- "Overpotential study:Relative charge flow between parallel Pt-Pt and Au-Au electrode systems in electrically isolated saline solutions".
- "Measurement of Electrode Bulk Resistance".
- "Bias Current Determination for a Current Source Biphasic Stimulator".
- "Electrode Scratch Study - Effects on Real Surface Area".
- Part of the protocol for the F.D.A. cochlea stimulation study.

Part 1
BACKGROUND

Summary: Part one of this thesis covers background material which will be useful as a foundation for further discussion of the research discussed in parts 2 & 3. This includes a brief introduction to the human visual system, basic models and psychophysics. A comprehensive survey of the literature on image measures is given, including the place of image quality and clutter measures in this context. An outline of the psychophysical experimental methods and analysis techniques, used in this thesis, is given in the last chapter of this section.

Chapter 1

Introduction

Summary: *This chapter provides an overview of the work described by this thesis and an introduction to certain aspects of image quality measures, where quality is defined in the context of this thesis and clutter metrics are integrated into this concept. A very brief and basic introduction to the human visual system is given with some basic models. The final section highlights the sections of work in this thesis that are original contributions.*

1.1 Overview

Electronically displayed images are becoming increasingly important as an interface between man and information systems. Lengthy periods of intense observation are no longer unusual. There is a growing awareness that electronic displays should be designed, so that an optimal match between the displayed image and the perceptual properties of the human visual system (HVS) can be achieved. The resulting demands placed on the display characteristics may vary greatly, depending on the task for which the displayed image is to be used and the ambient conditions. Image specifications that are optimal for achieving such a match are clearly not the same for a home TV, a radar signal monitor or an infra-red targeting image display. There is therefore a growing need for means of objective measurement of image quality, where "image quality" is used in a very broad sense and will be defined later, but includes any impact of image properties on human performance in relation to specified visual tasks.

The aim of this thesis is to consolidate and comment on the image measure literatures, to find through experiment the salient properties of electronically displayed real world complex imagery that impact on human performance for well specified visual tasks (of relevance to real-world issues), and from the data collected consider the appropriate application of image measures to this imagery to predict human performance.

1.1.1 Overview of Image Metrics

We are living in the digital age, with the widespread presence of digital technology. Imaging technology is also rapidly becoming digital, in all areas including medical, military and even television broadcasting is going digital. This introduces advantages, such as high quality reproduction and the possibility of image processing by computer. However, the quantity of data generated gives rise to associated problems of transmission and storage.

Image processing techniques, such as enhancement and restoration, are used to improve the *quality* of degraded images or maintain their quality despite degrading processes such as compression. However, when these image processing techniques are employed, the question arises as to how to evaluate the image processing algorithms. This implies that a method is needed to quantify the quality of the processed image. Therefore measures are required which characterise the salient properties of an image in a quantitative manner; *i.e.* the particular qualities of the image need to be measured.

In the literature, the term "image quality measure" or "metric"[1] defines the term "quality" ultimately with reference to the subjective judgement of an human observer [2]. Frequently,this judgement is applied to images only in an aesthetic sense, as for example in the viewing of television images (Van Dijk and Martens, 1997; Malo et al., 1997). In this thesis however, the research is more concerned with the usefulness or *utility* of an image for practiced interpreters, than with its aesthetic appeal for naive observers. This position was really adopted as a consequence of the psychophysical experiments described in this thesis, in particular Chapter 4, which showed that image quality is only meaningful in the context of a specific visual task. The term "quality" is also used in a different way, referring to particular characteristics of an image such as edge, texture or clutter[3] properties. Obviously the image measures must be defined with reference to the human observer to be useful in a practical way.

The image measures applied to clutter (clutter measures) have a literature (see section 2.6 in Chapter 2.) which is separate from the image quality measure literature, but is viewed in this thesis as a specific application of the same type of image measure as that defined for image quality. However, the definition of clutter measures usually requires an additional condition, in that they are normally specified with reference to a visual target.[4]. This may not always be apparent, since some clutter measures are used in effect to define texture in much the same way as image quality measures (see Chapter 2). Nevertheless, as will be shown later, the term "clutter" in practice refers to the difficulty in seeing a target, so it makes little sense to talk about clutter without reference to a target or target class. A good illustration of this point is the Waldman clutter metric used in Chapter 9.

[1] The difference between measure and metric will be discussed in Chapter 2.

[2] Even though image quality is sometimes measured in a purely objective way, if the image of interest is to be viewed by a human observer, then the image quality (and its measure) are referenced to the human observer's assessment

[3] Clutter is here defined as any structure in the image apart from the target, which masks the target or confuses the observer as to the location and/or class of the target. This is a major subject of investigation in chapters 2 and 7.

[4] Visual target acquisition is discussed in Part 2 of this thesis.

A review of the literature shows a modest effort (based on the number of publications) over the previous 20 or so years to develop quantitative image quality criteria. There was little effort until the mid 1970's, before advances in image processing technology. By the end of that decade, this effort began to wane. A short spurt of activity occurred in the mid eighties and then decayed to a low level. However, these efforts have met with only limited success. In many cases, the objective measure of image quality does not correlate with its subjective evaluation. Two major reasons contributing to this result are:

(i) No allowance has been made in the quality model for the response of the human visual system (or there is a lack of understanding of the same);

(ii) Misinformation or incomplete knowledge on the part of the person designing or selecting the metric, concerning the purposes for which the human observer is using the image(s).

With reference to item (i), it is known that subjective judgements of image quality depend on the purpose for which the image is to be used (Biberman, 1973; Briggs, 1980; Loo et al., 1984). For example, it may be expected therefore that the relative importance of the various physical image parameters should reflect this when assessing subjective image quality. Further, one may expect that the combination of physical image parameters that are optimal for an applied task, should depend on the task considered.

These matters make it questionable as to whether image quality could ever be expressed in a single general measure. For example, as a result, most authors have attempted to define a measure of a single aspect or a small number of aspects of image quality, often in the context of some specific tasks. However, there have been some attempts to quantify image quality in a single measure or metric, with limited success (these will be discussed in Chapter 2). Metz (Metz, 1977) classifies image quality measures into three basic types:

(i) Physically measurable functions that describe a single aspect of the imaging properties of the imaging system, such as resolution (point spread function [PSF], line spread function [LSF], optical transfer function [OTF], modulation transfer function [MTF] (see Chapter 2) or noise (autocovariance function [acvf] or Wiener spectrum);

(ii) Numbers or functions that attempt to provide a correlate with image quality by combining some of the above descriptors or their derivatives, and, in some cases, by attempting to take into account the object to be imaged, *e.g.*, information capacity, various signal to noise ratios (S/Ns), expected squared difference between object and image *etc*, and

(iii) measures of performance of a human observer who employs the visual data provided by a series of images to make decisions. These empirical measures may be functions, such as detection curves or receiver operating characteristic (ROC) curves (see Chapter 3), or numbers derived from these.

According to Metz, if one defines high quality images as ones that are *useful*, then these are appropriately characterised by type (iii) descriptors. Only type (i) measures are immediately

available to characterise an imaging system. Although type (ii) measures attempt to relate type (i) and (iii) measures, as discussed in Chapter 2 and shown in this thesis, they are unreliable and often disagree due to the dependence of the measure upon the application.

A useful tool to bridge between the physical measures and the human observer performance is signal detection theory (see Chapter 3). Firstly, it can be used to compute type (iii) measures of the performance of a hypothetical "ideal" decision maker, operating on physical image data which are similar to those available to a human observer (Green and Swets, 1966b; Barrett et al., 1993; Burgess, 1995); these image data are obtained through the application of the type (i) measures of the system as well as information about the imaged object. By using the ideal observer as a first order model of the human observer, it is possible to predict the effects of various object and imaging system properties on decision performance. It is also possible to attack the problem from the other direction, by inferring type (i) image attributes from empirical type (iii) measures of human observer performance, combined with a knowledge of the visual properties of the human observer. This latter approach is adopted in this thesis, where both a knowledge of the performance, and the visual properties, of the human observer were obtained through experiment.

1.2 The Human Visual System (HVS)

The following is a brief function description of the HVS. This description gives a basic account of the structure, function and psychophysics of the HVS. Some basic models of the HVS are also presented.

1.2.1 Functional Description of the HVS

Shown in figure 1.1 is a diagram of the cross-section of the eye, with the major components labelled. Light enters the eye through the cornea and lens which together focus this light as an image on to the retina, at the back of the eye. The amount of refraction of the light is controlled by the thickness of the lens, which in turn is controlled by the Ciliary muscles attached to it, thus allowing focusing of the image. The eye optics are not perfect and have a spread function that degrades the image. Further image degradation is caused by involuntary eye movements. These consist of slow drifts from the point of fixation, that occur at intervals of about 0.3 to 0.7 seconds, as well as high frequency tremors. Though these movements degrade the image, they are important in maintaining continuous visibility of the visual field, since an image that is stationary on the retina fades and eventually disappears. These eye movements are distinct from rapid eye movements known as *saccades* which direct the gaze to a point of fixation in a scene.

The retina itself is a complex structure, where a considerable amount of image pre-processing occurs. The retina is shown diagrammatically in figure 1.2. The retina consists of several different layers of different cells. The light-sensitive layer of the retina consists of photoreceptors, which are at a point of the layer farthest from the centre of the eye. Therefore, light must pass

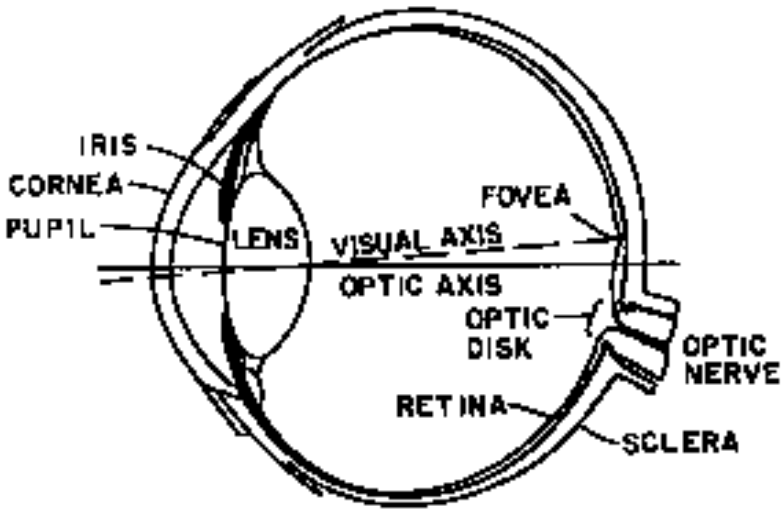


Figure 1.1: Cross-Section of the eye, showing gross anatomy. (From "Digital Pictures - Representation and Compression" by A. N. Netrauali and B. G. Haskell, 1988, Plenum Press.)

through all the layers of nerve cells before reaching the receptors.

The photoreceptors, which contain photosensitive pigments, are of two kinds: rods and cones. In the region surrounding the fovea, only densely packed cones are found. As the distance from the fovea is increased however, the density of cones decreases while the rod density increases. Cones are primarily responsible for spatial acuity and colour vision at normal daylight levels (*photopic* vision), while rods facilitate low light monochromatic (*scotopic*) vision. There is a light level range between the photopic and scotopic ranges known as the *mesopic* region where both rods and cones provide vision. These photoreceptors become less sensitive as the ambient light level increases, that is they *adapt* to different light levels. This is achieved by the light bleaching the light sensitive pigment in the photoreceptors, thereby reducing their light sensitivity. A new dynamic equilibrium is set up for each light level.

Shown in figure 1.2 is the simplified diagram of the interconnection of the various cells in the retina. As can be seen, the photoreceptors synapse (chemical connection) with the bipolar cells that populate layers of the retina closer to the lens. These bipolar cells in turn synapse with the ganglion cells, whose axons form the fibres of the optic nerve, along which the image data is transmitted (in the form of electrical impulses) to the brain. A small area, around where the optic nerve leaves the eye, is devoid of photoreceptors, resulting in a "blind spot".

Numerous lateral connections exist between the neurons in the retina, such as horizontal and amacrine cells. These are responsible for amplitude companding[5] and spatial frequency pre-emphasis of the visual signal by mediating the sensitivity of the ganglion cells to light. This *lateral inhibition* results in a reduction of signal from a cell when the neighbouring cells are also producing a signal. Lateral inhibition is mediated by the lateral connections of the horizontal and

[5] *C*ompanding is short for compressing and expanding. This is used in engineering systems to improve dynamic range, where the compression is usually a log-like function and expansion is via the inverse function.

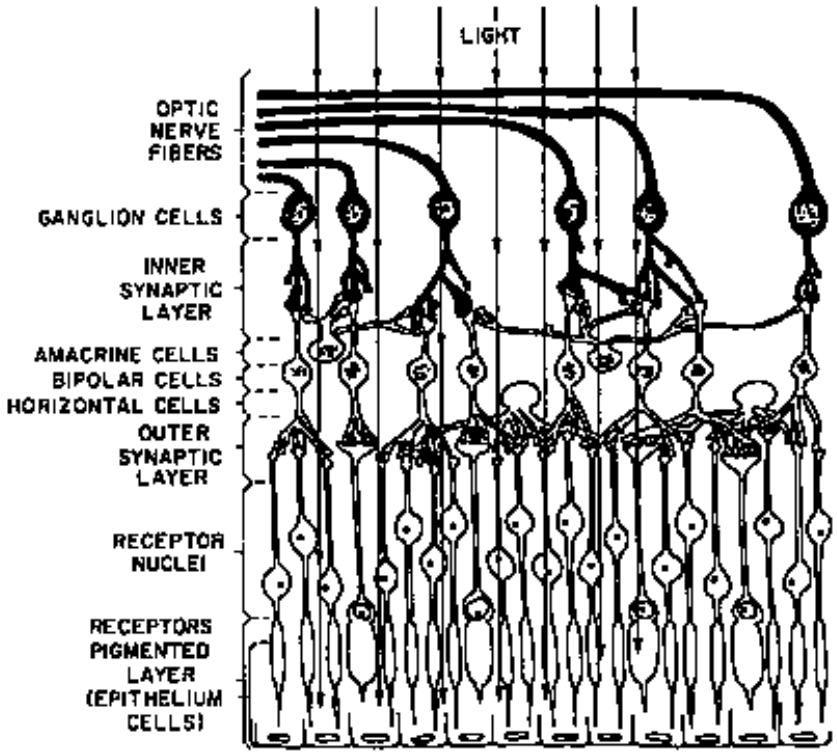


Figure 1.2: Structure of the Retina. Representation of the interconnections between receptors and bipolar, ganglion, horizontal and amacrine cells. (From "Organization of the Primate Retina: Electron Microscopy", by J. E. Dowling and B. B. Boycott, Proc. Royal Soc. B, 166, pp. 80-111)

amacrine cells. This results in a *receptive field* for each ganglion cell that exhibits an excitory region in the centre surrounded by an inhibitory region; *i.e.* light stimulating a ganglion cell raises its response, but light stimulating surrounding ganglion cells inhibits its response (the total response being the sum). Thus a stimulating pattern with "centre on surround off" will elicit the greatest response. These receptive fields are generally circularly symmetric.

The nature of transmission of visual information is by electrical impulses (action potentials). At synapses these signals are transmitted via chemical agents, over a small gap, which modify the potential at the target cell membrane. The signal is encoded by the "firing rate" of the transmitting cell.

Shown in figure 1.3 is a plot of firing rate versus stimulus intensity, for various levels of light adaptation. The active response region is shifted with different background illumination levels. In the active region the firing rate is almost linearly related to the log of the intensity and virtually independent of background illumination.

Figure 1.4 shows a schematic diagram of the visual pathway beyond the retina. Most of the visual processing takes place at this stage, but is not completely understood and beyond the scope of this thesis; however, a basic description of the visual pathway follows. At the Optic Chiasma the fibres of the optic nerve split in a way that depends upon which half of each retina

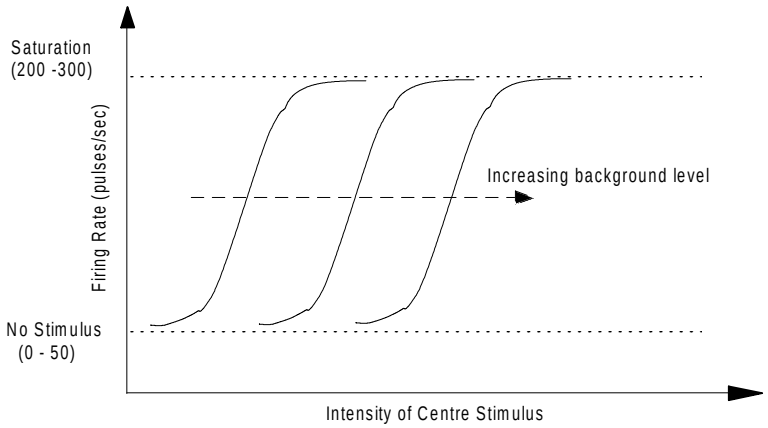


Figure 1.3: Light Level Adaptation: Firing rate as a function of stimulus intensity for several background intensity levels. (From "Digital Pictures - Representation and Compression" by A. N. Netrauali and B. G. Haskell, 1988, Plenum Press.)

they originate in. Fibres from the right half of each retina go to the Lateral Geniculate Nucleus (LGN) on the right side of the brain and fibres from each left half retina reach the LGN on the left side of the brain. At the LGN each half of the visual field is mapped to the appropriate part of the visual cortex, where significant visual processing and visual perception occurs.

1.2.2 Visual Psychophysics

In visual psychophysics the "black box" approach is taken, where visual stimuli are the inputs and prescribed sensations are the output. The "transfer function" of the black box describes the system. In the context of image quality, the consideration of just noticeable differences (JNDs) is important. In relation to this, an important parameter is the *visibility threshold* of a stimulus. This is defined as the magnitude of a stimulus that becomes just noticeably different; *i.e.* the probability of detection for a human observer is 50%. There has been considerable work done in the past on this subject, which will be briefly discussed. However, there are fewer data on suprathreshold psychophysics, probably due to this being a much more difficult and less precisely defined problem.

1.2.3 Models of the Human Visual System

During the past few decades it has become common to model the HVS. This is probably due to the recent increase in physiological and psychophysical data, coupled with technological advances. With the advent of digital image processing and analysis, and more recently computer vision, impetus has been given to visual modelling.

In earlier times, the aim of visual models was to explain the functioning of biological systems without regard to applications. However, recent models are frequently intended to be integrated into hardware and software systems and have thus become more mathematical in nature.

Here only monochrome models are discussed, but colour models are usually a straight

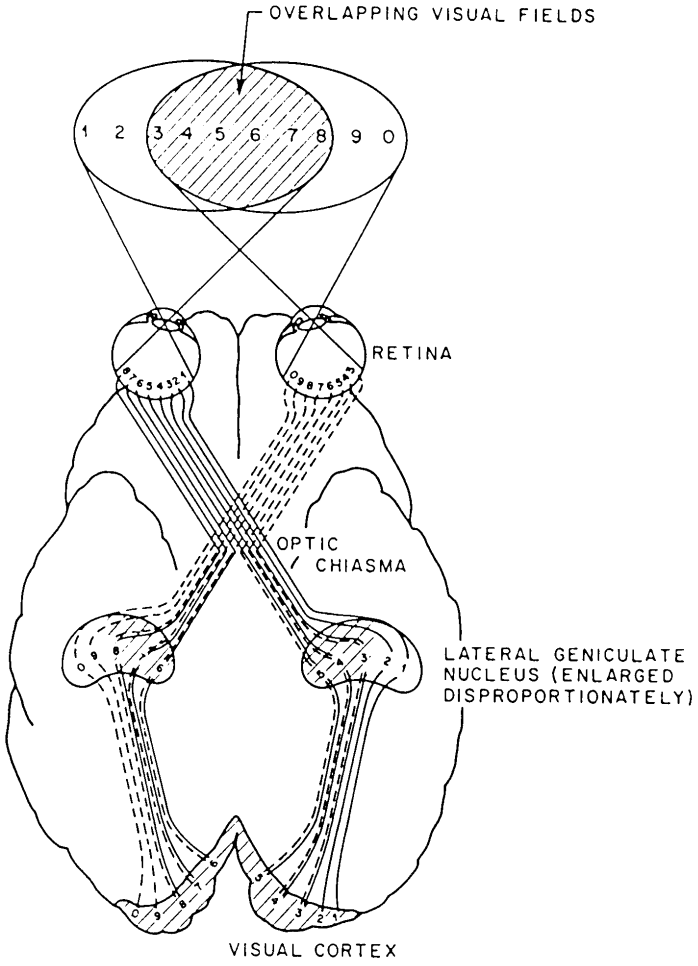
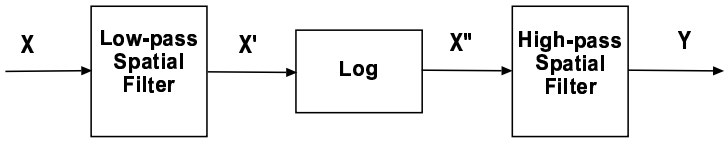


Figure 1.4: Visual Pathway: Diagram of the visual pathways from each eye to the visual cortex, via the Optic Chiasma and the Lateral Geniculate Nucleus (LGN). (From "Digital Pictures - Representation and Compression" by A. N. Netrauali and B. G. Haskell, 1988, Plenum Press.)

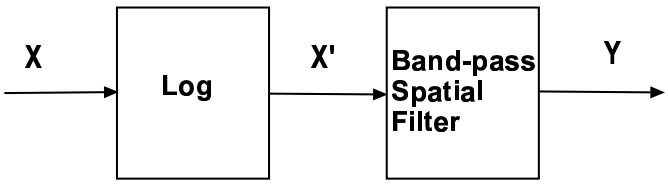
forward extension involving three spectrally dependent versions of the monochrome model. Described here are some of the early and basic mathematical models which address the salient properties of the HVS. There are many more complicated and complete models which are covered quite well in a book edited by Peli (Peli, 1995).

Shown in figure 1.5(a) is a simple achromatic model of the HVS. This model was found to have a high frequency roll-off that was a function of contrast. In particular, the system sensitivity to high spatial frequencies was found to decrease with increase in contrast (Hall and Hall, 1977). However, this model can be simplified, as shown in figure 1.5(b), and does not exhibit the above properties (Hall, 1981). It is assumed in this simplified model that the input intensity range is in a linear portion of the logarithmic curve. This allows the low-pass spatial filter to be combined with the high-pass filter to produce a band-pass filter. Hall (1977) used filters depicted in figure 1.5(a) with the following transfer functions:

$$H_{LP}(\omega) = \frac{0.14}{0.49 + \omega^2} \tag{1.1}$$



(a) Achromatic HVS Model



(b) Further Simplified HVS Model

Figure 1.5: A simple achromatic model of the HVS produced by Hall (1981). Here the non-linear intensity response of the HVS is modelled by the logarithmic (log) function, while the spatial frequency response of the HVS is modelled by simple filters.

This corresponds to 3mm diameter pupil and has its 3dB cutoff at 6.6 cycles/degree.

$$H_{HP}(\omega) = \frac{10^{-4} + \omega^2}{4x10^{-3} + 0.8\omega^2} \tag{1.2}$$

1.2.4 Contrast Sensitivity Model

Consider an intensity pattern $u(x, y)$ of a grating at an angle θ to the horizontal, so that

$$u(x, y) = U + \alpha p(x \cos\theta - y \sin\theta), \tag{1.3}$$

where U is the background level and $p(\cdot)$ is a periodic function (quite often a sinusoid).

For many variations of p and U researchers have measured the contrast sensitivity; *i.e.* the ratio $\frac{U}{\alpha}$ at which the subject can just detect the grating with a uniform background. Manos and Sakrison (1974) state, based on the work of others, that the contrast sensitivity model has the form

$$\left(\frac{U}{\alpha}\right)_{\text{threshold}} \approx c\frac{f}{f_0}e^{-\frac{f}{f_0}}. \tag{1.4}$$

The value of f_0, the peak of the curve (figure 1.6), falls between 3 and 5 cycles/degree of viewing angle. In this model the HVS is assumed to be isotropic, even though this is not true. After

the initial nonlinearity, the HVS response can be considered linear over a moderate range of intensities. The nonlinear component is described as a monotone increasing convex function of the form

$$f(u) = u^b, \tag{1.5}$$

where u is the pixel intensity and b is a positive real number.

The best value for b, according to subjective evaluation, was reported as being 0.33 (Mannos and Sakrison, 1974). Shown in figure 1.7 is the nonlinear intensity mapping function defined in (1.5), while figure 1.6 shows a plot of the linear part of the HVS model and is defined by

$$A(f_r) = [c_1 + c_2 \frac{f_r}{f_0}^{k_1}] e^{-\frac{f_r}{f_0}^{k_2}}. \tag{1.6}$$

The parameters c_1, c_2, k_1, k_2 and f_0 are chosen experimentally for best results. Typical values for the HVS are: $c_1 = 0.2$, $c_1 = 0.081$, $k_1 = 1.0$, $k_2 = 1.0$ and $f_0 = 5.55$ (Nill, 1985). The plot of (1.6) with the parameter values just listed, is shown in figure 1.6. Since (1.6) is assumed isotropic, the radial frequency is defined as $f_r = \sqrt{f_x^2 + f_y^2}$, where f_x and f_x are the spatial frequencies in the x and y directions respectively. The input to the linear component is the output of the nonlinear portion of the model. The type of models just discussed are useful

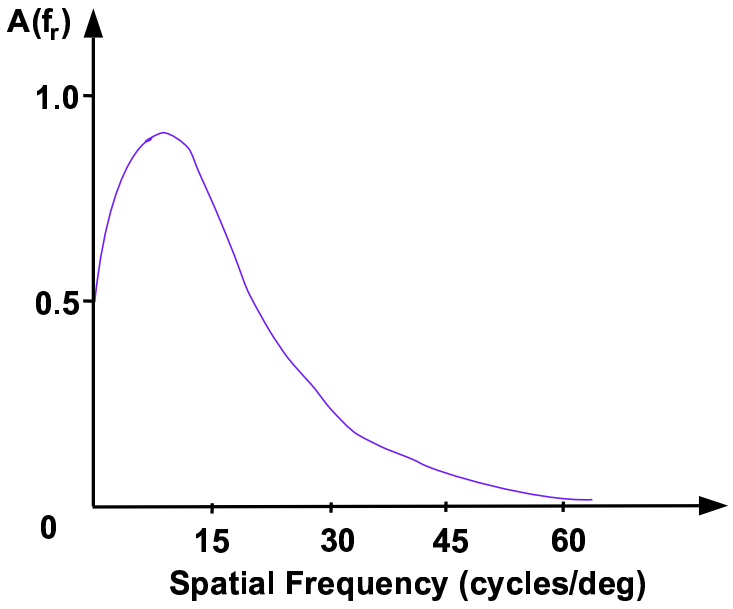


Figure 1.6: Model of HVS Spatial Response.

in describing the general behaviour of the HVS, but they are not adaptive to different types of image scenes being presented, nor do they reflect any "higher level" input. For example, in the context of a specific visual task, such as searching a scene for a specific target, the HVS "parameters" are modified by higher level input. The HVS is tuned to recognise the specific target, so knowledge about likely areas in the scene to search are included as input to the HVS.

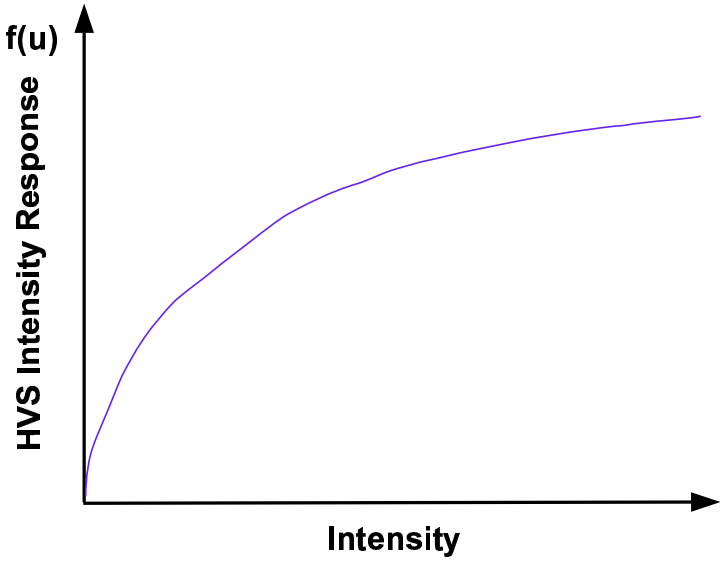


Figure 1.7: Model of HVS nonlinear Intensity Response.

1.3 Categories of Image Measures

Different approaches have been taken by researchers in categorising image measures, and one possible partition has been already discussed in section 1.1. Other authors have taken different approaches in categorising image measures, based on their mode of application, which are discussed in the remainder of this section and in Chapter 2. In terms of application, image measures have been divided into two main groups of either similarity measures or interpretability measures. Each of these can be evaluated by objective or subjective means at a global or local level (figure 1.8). It should be noted that these classification schemes are not mutually exclusive; *i.e.* they are not competing schemes.

1.3.1 Image Similarity and Image Interpretability

It has been stated in the literature that an important property of an image quality measure is that it correlates with image interpretability (or utility) (Briggs, 1980; Todd-Pokropek, 1976; Metz et al., 1976). However, image similarity (or fidelity) is important when the aesthetics of the imagery is paramount, such as in images for entertainment; *e.g.*, television viewing. In this thesis more consideration is being given to the former class of image metrics.

1.3.2 Global vs Local Image Measures

The application of image measures can be at the global or local level. Of course the definition of some measures may preclude the application of the measure at a particular level. In the case of a global measure the support for the measure is the entire image(s), whereas for a local measure the region of support is confined to some local area of interest.

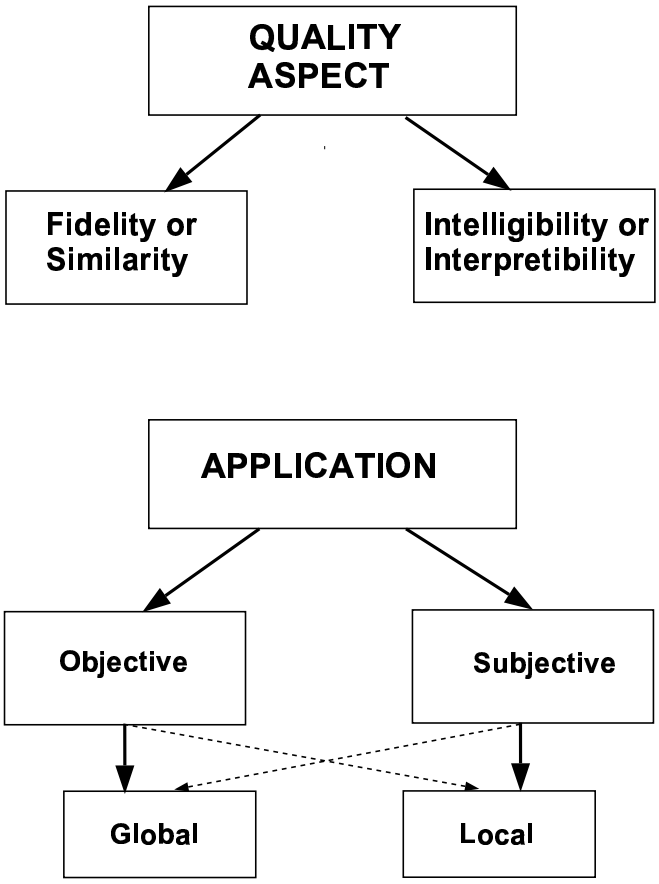


Figure 1.8: Categorisation of Image Measures.

In some instances (maybe most) a global measure may not correlate with an evaluation of the quality of an image by a user as he/she will tend to pay attention to those features in the image of relevance to his/her application. It has also been suggested that such a user will assess the image quality by judging the areas of most degradation; in fact it has been suggested that the user rates the image by a weighted average of the worst two or three regions (Limb, 1979; Ohtsuka et al., 1988).

Image measures that have been designed as global measures can sometimes be applied in a local sense. For example, they can be applied to the local neighbourhood of each pixel, in a moving window fashion, or to adjacent blocks (which is cheaper computationally). Following the suggestion of Limb cited above, the (2 or 3) neighbourhoods/blocks with the worst quality can be averaged to produce a measure of global image quality. It is interesting to observe that, if the image distortion is additive, then a local measure should be similar to the global measure and should approach it as the size of the region of support increases to include that of the whole image.

In the case where the interpreter/user is interested in some feature(s) in the image another approach is required. Here the local image measures must characterise the local feature. Examples of such features are edges and texture.

1.4 Contributions of this Thesis.

This section briefly outlines each chapter and identifies the parts I believe to be original contributions.

Chapters 1 and 2 provide an introduction to certain aspects of image quality measures, where "quality" is defined in the context of this thesis and clutter metrics are integrated into this concept. An analysis of image measures that have been classified according to the features they are attempting to quantify is provided. The relevant literature is reviewed and image measures are categorised according to their underlying principles and the intended mode of application. This provides new insights into the various image quality metrics - including clutter metrics - and their interrelationships.

Chapter 4 describes a study which investigates how humans relate to the same scene information presented both as infra-red (IR) imagery and optical imagery. Also a subjective methodology to produce an interval scale as a metric of image similarity was investigated, and some basic image quality metrics were implemented to get a feel for their application. Though this study was only preliminary, some new results were reported.

In the context of the interpretation of image similarity, it was found that the usual statistical type of image measure was inadequate. It was found that, in order to capture the complexity of the images, measures of local (region-based) image properties are required. In order to analyse the relationships between image objects, syntactic/semantic[6] (Gonzalez and Wintz, 1987a) type of measures are also required.

The aims of Chapter 5 include that of analysing human visual performance under a well-defined visual task, with imagery that has undergone degradation. Image compression was chosen as the means of degradation because it is controllable and image compression has practical application. Two methods of static image compression - JPEG and a fractal-based method - were compared in terms of the detectability of simple targets following compression and decompression of the images containing such targets.

This work presents new results on the comparative performance of JPEG and Fractal image compression. This is also novel in that compression algorithms, such as JPEG have traditionally only been tested by using subjective panels to rate the aesthetic image quality of the compressed images. In contrast, this study introduces the idea of image quality measures that are task oriented. The results of this work also give some useful rules of thumb for the use of image compression in the surveillance context. Further insights into human perception which have general application are also given.

Chapter 6 describes work on the effects of video compression on visual tasks. This work continues the still image compression experiments on target detection on in a similar vein, but examines the effects of video compression on target recognition. A set of video compression experiments were performed which required observers to recognise targets in randomly presented video sequences. The sequences had controlled levels of contrast and multiplicative noise, and

[6] That is, structural and relational aspects of the objects within the image scene itself.

were compressed and de-compressed at a variety of compression levels using MPEG-2 encoding under standard settings.

Again, as for still compression, the usual protocol for testing video compression, particularly by MPEG working groups, was to do subjective rating of aesthetic video quality. This study took a new approach in analysing the effects of MPEG compression on human visual performance. This would be important if, as is likely, compression is used in a surveillance or similar context. The effects of MPEG-2 compression in this context were quantified and it was found that large compression ratios were required to significantly degrade human performance. The experiment indicated that this was largely due to temporal integration in the HVS. Learning effects were examined and quantified and comments made on individual differences in perceptual cognition.

Chapter 7 deals with the effects of clutter on human target acquisition. In theory, properties of clutter can be defined globally or locally. However, in the literature, the distinction between local and global clutter is arbitrary, and the standard approach of setting the local domain (support of the metric) to twice the expected target size is adopted without any justification. This work addresses this issue and considers the implications for the application of clutter metrics.

Although this study is largely exploratory, it has shown that the accepted practice in applying clutter metrics is incorrect. The local extent to which clutter effects target detection was shown to be, in this instance, much greater than twice the target size, for targets smaller than 0.8° radius. A model was presented explaining these phenomena, indicating that the auto-covariance function characterising the clutter is the main determinant of the size of the region of local clutter, and that this region is reduced for larger targets.

Chapter 8 considers the effects of image processing parameters on the clinical value of Single Photon Emission Computed Tomography (SPECT) images. The overall long-term aim is to develop an automatic system for optimal image filter parameter adjustment.

A new measure, called the gradient energy measure (GEM), for quantifying the effect of filtering on SPECT images was developed and evaluated. This proved to be a reliable measure of image smoothing and noise level, which, in preliminary studies, agreed with human perception. There is a model of HVS function that appears to be modelled to a first order by the GEM. This model is known as the "energy integrator" model (Green and Swets, 1966c; Moulden et al., 1990).

Chapter 9 describes a study which determined the performance of human image analysts in the surveillance context, using Synthetic Aperture Radar (SAR) derived images, in terms of the analyst's receiver operating characteristic. The experiment was designed to correspond as closely as possible to the expected real world mode of operation of the analysts using similar imagery.

The effects of target contrast and background clutter on human analyst target detection performance were quantified and the Waldman (Waldman et al., 1988) clutter metric was validated with real imagery. The findings in Chapter 7 were also re-enforced with real data in a pseudo-operational context. The signal-detection-theory paradigm was extended by using the

ROC in a non-classical way, and parametric versus non-parametric ROC development was explored, showing the latter to be more robust in this application. Many issues in regard to setting up and performing a complex real world experiment were explored and discussed.

Chapter 10 contains a summary of the results of the thesis. In addition, the section on further longer term work presents a proposal, based on the research in this thesis, for a new system for performing image quality assessment; *i.e.* a system for predicting, given an actual digital image, the human evaluation of image quality and utility.

Chapter 2

Objective Measures of Image Properties

Summary: *This chapter contains an analysis of image measures that have been classified according to those features they are attempting to quantify. The relevant literature was reviewed and image measures were categorised according to their underlying principles and their intended mode of application. Many of these measures can be applied locally or globally, but some, such as edge measures, are local in nature. These measures can also be classified as either similarity (fidelity) or interpretability (intelligibility) measures. Clutter measures are regarded as a form of the latter. Note that these classification schemes are mutually inclusive.*

2.1 Introduction

The introductory chapter gave an overview of image measures and some broad classification of these measures based on their purpose and mode of application; *e.g.*, to be applied locally or globally, and/or to measure image similarity or image information (interpretability). This chapter contains an analysis of image measures which have been classified according to the features they are attempting to quantify. In scrutinising the literature I have found five basic classes of image measures to present themselves. They are listed here and will be discussed in detail in the rest of this chapter.

(i) Distance measures, the most common being the mean square error (MSE);

(ii) Modulation transfer function (MTF) type measures. These are commonly used in assessing imaging systems but can be used in assessing images directly;

(iii) Information theoretic measures;

(iv) Decision theoretic measures;

(v) Signal detection theoretic measures.

The above listed measures are usually applied in a global sense, but could be applied with local support; *i.e.* calculated over a localised area of the image. There exist another two classes of measures that are usually applied to local features. These are "edge quality measures" and "texture measures", which are discussed in section 2.3, where often texture measures are in the class of the entropy based measures; *i.e.* they are often an information theoretic measure (iii).

Loo, Doi and Metz (1984) classified image quality measures into two broad categories:

(i) displayed models: these treat the human visual system (HVS) as a perfect transfer device; and

(ii) perceived models: these include the image transfer characteristics of the eye-brain system.

2.2 Global Objective Measures

As outlined in Chapter 1, the application of image measures can be at the global or local level. Of course, the definition of some measures may preclude the application of the measure at a particular level. In the case of a global measure the support for the measure is the entire image(s), whereas for a local measure the region of support is confined to some local area of interest. This section addresses the application of measures in a global sense, some of which may also be applied in a local context. The next section addresses specifically local measures.

2.2.1 Distance and Related Measures

Measures aimed at rating image similarity are necessarily bivariate in nature; *i.e.* they require two images to compute the measure. Here an image, of which the quality is to be judged, is compared to some reference image. A typical application of this idea is where an image is compressed and sent down a communications channel and then reconstituted, with the pre-compressed image as the reference image to which the transmitted image is compared.

Now the quality of the image in question is determined by how closely it matches the reference image, which can be obtained by measuring the difference or "distance", in some space, between the object image and the reference image. In the literature several approaches have been adopted to measure these differences, but the main technique has been the use of the p-norm distance measure, which is now defined.

Let $\hat{X}$ be an image that is to be compared to an image X so as to ascertain its quality. Let both X and $\hat{X}$ consist of pixels with position co-ordinates (i, j) with $i, j \in \{1, \ldots, N\}$. Further, let the intensity values of the pixels of X and $\hat{X}$ be given respectively by x_{ij} and $\hat{x}_{ij}$, where it is assumed $\forall\, i, j : x_{ij}, \hat{x}_{ij} \in \{0, \ldots, a-1\}$, with a the number of intensity levels.

Definition 2.2.1

The p-norm distance measure for the difference between X and $\hat{X}$ is

$$d^p(X,\hat{X}) = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} |x_{ij} - \hat{x}_{ij}|^p \right\}^{1/p} \tag{2.1}$$

where $p \geq 1$. From the previous definition the following properties for $d^p(X,\hat{X})$ can be derived.

Corollary 2.2.1

$$d^p(X,\hat{X}) = 0 \text{ iff } X = \hat{X}, \tag{2.2}$$
$$d^p(X,\hat{X}) = d^p(\hat{X},X), \tag{2.3}$$
$$d^p(\hat{X},\check{X}) \leq d^p(\hat{X},X) + d^p(X,\check{X}). \tag{2.4}$$

The properties listed above define what is called a "*metric*" (Rosenfeld and Kak, 1982). Obviously from property 2.2, the distance measure must achieve its minimum when the object image is the least distorted. Property 2.3 is trivial and self explanatory. Property 2.4 is a result of the Minkowsky inequality (Mitrinovic, 1970).

There exists some commonly used special cases of the class of distance measure shown in definition 2.2.1. These include the following:

$$d^1(X,\hat{X}) = \sum_{i=1}^{N} \sum_{j=1}^{N} |x_{ij} - \hat{x}_{ij}|. \tag{2.5}$$

This is the total absolute difference or error. Another measure is the peak-error (Rosenfeld and Kak, 1982) which is defined as

$$\lim_{p\to\infty} d^p(X,\hat{X}) = \max_{ij} |x_{ij} - \hat{x}_{ij}| \equiv d^\infty(X,\hat{X}), \tag{2.6}$$

where the Euclidean distance is defined as

$$d^2(X,\hat{X}) = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} (x_{ij} - \hat{x}_{ij})^2 \right\}^{1/2}. \tag{2.7}$$

One shared failing of these measures is that they are dependent on the number of pixels in the region of support. To overcome this difficulty, these measures can be modified to reflect the expected value per pixel. Thus, for example, the distance measure expressed in equation 2.7 can be modified as follows to express a very commonly used measure, the mean square error (MSE); *viz.*

$$MSE = \left\langle (x_{ij} - \hat{x}_{ij})^2 \right\rangle \text{ for } i,j = 1,\ldots,N \tag{2.8}$$

where $\langle\cdot\rangle$ is the mean or expectation operator.
A somewhat less used measure, more closely related to equation 2.7, is the root mean square (RMS) measure; viz.

$$RMS = \left\{\left\langle (x_{ij} - \hat{x}_{ij})^2 \right\rangle\right\}^{1/2} \text{ for } i, j = 1, \ldots, N \tag{2.9}$$

The MSE is more sensitive than the RMS to large but localised deviations between X and $\hat{X}$, since this applies in general for $p > q$ with $d^p(X, \hat{X})$ and $d^q(X, \hat{X})$ analogous to MSE and RMS respectively.

Definition 2.2.2
The MSE measure can be generalised as follows:

$$d_m^p(X, \hat{X}) = \{\langle |x_{ij} - \hat{x}_{ij}|^p \rangle\}^{1/p} \text{ for } i, j = 1, \ldots, N \tag{2.10}$$

for $p \geq 1$.
Note: d_m^p is equivalent to the exponential mean (Korovkin, 1961).

The selection of a particular measure depends upon the particular emphasis that is required for the situation in which it is being used; *e.g.*, (2.6) would be used if the peak error was of most interest *etc.* It can be shown that the influence of large differences between X and $\hat{X}$ becomes greater for increasing values of p. With $p = 1$, all differences are weighted equally, but as $p \to \infty$ larger differences are weighted progressively more until only the maximal difference of the entire image is effectively measured (2.6). As shown in section 1.2 of Chapter 1, large values of p give a measure that corresponds more with a property of the HVS, in that the overall perceived image quality is less affected by small differences than by larger ones. The previous discussion on the effects of increasing p are further supported by an interesting corollary (Beckenbach and Bellman, 1971).

Corollary 2.2.2
If $p_1 > p_2 \geq 1$ then

$$d^{p_1}(X, \hat{X}) \leq d^{p_2}(X, \hat{X}) \tag{2.11}$$

This corollary shows that $d^p(X, \hat{X})$ is a decreasing function of p, whereas $d_m^p(X, \hat{X})$ can be shown (Korovkin, 1961) to be an increasing function of p (corollary 2.2.3). Notwithstanding this it is easily shown that $d_m^p(X, \hat{X})$ (definition 2.2.2) is a metric as defined in corollary 2.2.1. There is however a difference in the behaviour of $d_m^p(X, \hat{X})$, as p increases, compared to that of $d^p(X, \hat{X})$.

Corollary 2.2.3
If $p_1 > p_2 \geq 1$ then

$$d_m^{p_1}(X, \hat{X}) \geq d_m^{p_2}(X, \hat{X}) \tag{2.12}$$

The bounds of the $d_m^p(X, \hat{X})$ measures are set by the following:

Corollary 2.2.4
Let the images X and $\hat{X}$ be digitised into a maximum of a levels; *i.e.* let $x_{ij}, \hat{x}_{ij} \in \{0, \ldots, a-1\}$, then:

$$\min_{ij} d_m^p(X, \hat{X}) = 0, \tag{2.13}$$

$$\max_{ij} d_m^p(X, \hat{X}) = (a-1). \tag{2.14}$$

The absolute maximum shown in (2.14) is only achieved if the reference image and the object image contain pixels with values of 0 and $a-1$. In fact to achieve this absolute maximum each pair of corresponding pixels would have to have a difference of $(a-1)$, since $d_m^p(X, \hat{X})$ measures the distance per pixel. Therefore in practice only sub-maximal values are achieved.

2.2.2 Measures Incorporating Decision Theory

In order to address the problem of finding an image measure that allows for the characteristics of each individual image under consideration, a decision theoretical based measure was introduced by Spaulding (Spaulding and Engeldrum, 1985). Decision theoretical concepts will be introduced as needed in this section. The work described here is in the context of monochromatic photographic tone reproduction, which has been classically characterised by the "D-log-E" curve. This curve relates the optical density[1] of the photographic image to the logarithm of the exposure. It is asserted that the optimum photographic system will produce an image that reproduces exactly the relative luminances in the original scene, where relative luminance means the luminance[2] of interest referred to the scene reference white luminance. This implies that the "D-log-E" curve is linear for an optimum system.

Spaulding's paper explores the quality of tone reproduction in photographic images. This paper gives a brief historical review of the work done in this area, which includes a discussion of the effects of surround illuminance on the perceived brightness. Included in this discussion is a measure for tone reproduction quality introduced by Bartleson (Bartleson, 1975), which is defined to be

$$\Delta = \sqrt{\sum (B_0 - B_i)^2}, \tag{2.15}$$

where Δ is defined as the deviation from optimum quality, B_i is the original image or scene brightness relative to original image or scene white, and B_0 is the reproduction brightness, relative to the reproduction white reference. The gradient of the brightness, relative to the reference white, versus log luminance of the reproduction system varies with surround brightness, so as to produce the same perceived gradient as that of the original image. Bartleson weighted the data to obtain a high degree of correlation; this is almost certainly due to the psychophysical

[1] See Appendix A for a definition.
[2] See appendix A for a definition.

response of the human visual system. Bartleson (Bartleson and Breneman, 1967) developed a simplified function to describe the subjective response to brightness, which is

$$L^{**} = A(100\frac{R}{R_0} + B)^{\alpha} - 16, \tag{2.16}$$

where $L^{**}$is the lightness (renamed to avoid confusion with earlier terminology), R is the reflectance, R_0 is reference white, and A, B and α are constants which depend on surround luminance. These tones for the original image and a reproduction image are determined with respect to a reference white. Dom (1981) applied this equation and found the expected linear relationship between the relative brightness of the scene and original images for optimum reproduction. He states that this result does not necessarily apply to scenes outside the scope of the (typical) set he used. However, Spaulding suggests that these scenes did cover a wide range of brightness distributions.

The main thrust of this approach is to apply concepts of decision theory to image measures, to allow for the brightness distributions that are found in different scenes. Now, some of these decision theoretical concepts will be introduced. The discrepancy between the scene and the reproduced image is defined in terms of the "expected loss" as

$$R[\theta, d(z)] = E\{L[\theta, d(z)]\}, \tag{2.17}$$

where $R[\cdot]$ is the the "risk" or expected loss, $L[\cdot]$ is the loss function or the actual metric to determine the degree of variation from optimal, θ is the "state of nature" or the actual reproduced relative brightnesses, and $d(z)$ is the "decision function". He associates B_0 with the reproduced relative brightness and lets $d(B_i) = B_i$; *i.e.* it is "decided" that the scene relative brightness be equal to the reproduced original relative brightness, which was shown earlier to produce the optimum result. Then

$$L[\theta, d(z)] = L[B_0, B_i] = (B_0 - B_i)^2. \tag{2.18}$$

Assuming B_i to be a random variable, Spaulding asserts that, for a given image I, the distribution of B_i can be characterised by the conditional probability $p(B_i|I)$. The expected loss for the reproduction system is defined as

$$R[B_o, B_i|I] = \int_0^{100} L[B_o, B_i]p(B_i|I)dB_i. \tag{2.19}$$

The value of $p(B_i|I)$ is determined from its approximation, the pixel histogram. Equation 2.16 was used as a measure of the relative brightness of a series of images, with suitable values substituted for the parameters. The limits of the integral in (2.19) reflect the range used for relative luminance or brightness in practice. Subjective responses for reproduction quality were obtained using a "magnitude estimation" method (see chapter 3). The measure used did not exhibit a linear relationship with subjective responses as expected, so a more general power function was incorporated into the loss function, yielding the following definition

$$L[B_0, B_i] = |B_0 - B_i|^N. \tag{2.20}$$

The value of N was tested for values from 0.1 to 5.0., with N=3 giving the best correlation of 0.73, using a log-log plot. (The model was asserted to be: $Y = AX^p e^{\varepsilon}$, which transforms to $\log Y = \log A + p \log X + \varepsilon$ in the log domain). The authors admit the inadequacy of these measures, suggesting that the subjective process has not been modelled appropriately. They suggest that the assumption that all areas of the scene have equal importance is incorrect. This seems to me to be likely, not only at the early visual level, but also the higher levels which are driven by task related goals.

2.2.3 Measures Incorporating Signal Detection Theory

The approach to image evaluation described here is predicated upon measuring the effects of image properties on the decisions made by an observer using an imaging technique in a given situation. The approach is divided into two parts:

(i) Measurement of the relationships among the relative frequencies of the various types of correct and incorrect decisions made by an observer; and

(ii) evaluation of the benefit to be gained from those possible combinations of decision frequencies.

Principles of signal detection theory are used to guide the approach, to predict the relationships among descriptors of decision performance in various situations and to suggest optimal decision making strategies.

The basic principles of the *receiver operating characteristic* (ROC), which is under-pinned by signal detection theory, are discussed in chapter 3 on page 61.

2.2.4 MTF Based Measures

The optical transfer function (OTF), which will be defined shortly, was developed earlier this century to define optical system performance. The OTF, used with an appropriate measure of gross scene contrast, specifies completely an optical image. However, there is no direct relationship between OTF and visual performance.

Spatial Frequency Response

If the optical system is linear, then its characteristics can be determined from its spatial frequency response. This is a 2-D concept, but it is usually tested and analysed in terms of two orthogonal 1-D bar patterns; *i.e.*

$$B_L = B_m + B \cos 2\pi f_s x, \tag{2.21}$$

where B_L is the local luminance, B_m is the mean luminance, B is the maximum deviation from the mean, f_s is the spatial frequency and x is the distance from the origin. Then, applying

(2.21) to the input of a general linear system will result in the output

$$B'_L = B_m + \gamma \cos(2\pi f_s x + \phi), \tag{2.22}$$

where γ is the relative amplitude and ϕ is the phase angle. Then $\gamma(f_s)$ is called the Modulation Transfer Function (MTF), while $\phi(f_s)$ is called the Phase Transfer Function (PTF). The combination of MTF and PTF is called the Optical Transfer Function; *i.e.*

$$OTF = F(j\omega) = \gamma e^{j\omega x}, \tag{2.23}$$

where $\omega = 2\pi f_s$. The OTF is related to the point spread function via the Fourier transform

$$F(j\omega) = \int_{\infty}^{-\infty} G(x) e^{-j\omega x} dx, \tag{2.24}$$

where G(x) = line spread function.

Point Spread Function

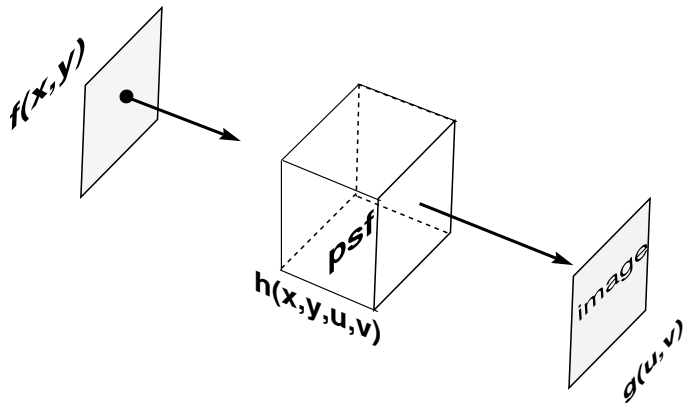


Figure 2.1: A simple representation of the effect of a PSF on image formation.

Definition 2.2.3
Let f(x,y) be the input to an optical system with system function h(x,y,u,v) which gives output g(u,v). If the PSF is invariant under translation then h(x,y,u,v) has the form h(u-x,v-y); *i.e.*

$$g(u,v) = \int_x \int_y f(x,y) h(u-x, v-y) dx dy \tag{2.25}$$

The PSF is obviously the 2-D transfer function; *i.e.* the output is the convolution of the input and the PSF.

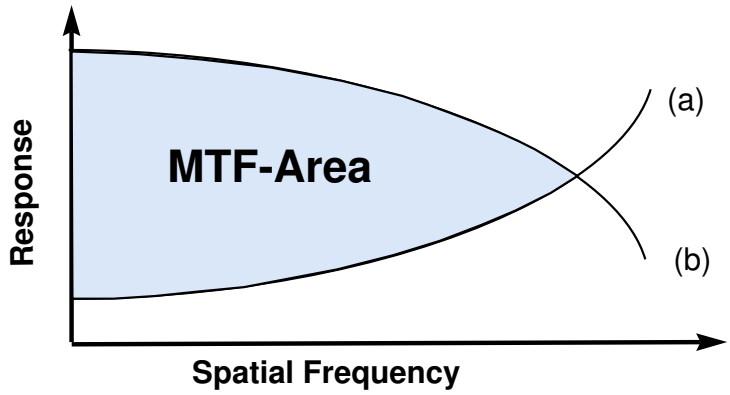
The OTF characterises the system (or frequency) response of the optical system which determines the final image quality, but does not give a measure of the human subjective visual response. To try to overcome this, other measures based on OTF, but incorporating knowledge of the human visual system, have been developed.

Modulation Transfer Function Area

Of the MTF-based measures, the Modulation Transfer Function Area (MTFA) is the most commonly used in practice. As illustrated in figure 2.2, the MTFA is defined as the area between the MTF curve of the optical system under consideration and the detection resolution threshold curve for the HVS, which is also known as the spatial contrast sensitivity function (CSF). The latter is usually measured under the optimal viewing of a standard resolution test object such as the American 3-bar test target. The MTFA is formally defined as

$$MTFA = \int_0^{\nu_0} [MTF(\nu) - CSF(\nu)]d\nu, \tag{2.26}$$

where ν is the spatial frequency in cycles/degree. By definition, the "crossover" frequency ν_0, is the frequency at which $MTF(\nu) = CSF(\nu)$; *i.e.* where the curves (a) and (b) in figure 2.2 intersect. The CSF is the approximate inverse of the retinal MTF for observer and viewing system combination. Thus the MTFA is approximately representative of the area under the system MTF referred to the retina. In 1973 Snyder found strong correlation between subjective ranking of goodness and MTFA for photographs. Later, Higgins (1977) found a stronger correlation using the square of the area under the resultant curve by multiplying the instrumental MTF and that of the eye refractive optics. Feng *et al.* (1990) found the MFTA to be a suitable measure of image quality on visual display units, primarily in the context of text displays. The MFTA was adopted as a standard by the USA in 1988 (ANSI/HFS, 1988; Snyder, 1989), However, Infante (1991) points out that the method for calculating the MFTA, given in this standard, is incorrect and describes what he argues is the correct procedure.



(a) HVS Threshold detectivity curve.
(b) MTF of display system.

Figure 2.2: This figure illustrates the calculation of the Modulation Transfer Function Area (MFTA). The MFTA is determined by calculating the area between curves (a) and (b) as shown by the shaded area.

Acutance Measures

These were developed to attempt to predict subjective "sharpness" of images (photographs). After some evolution, the Contrast Modulation Transfer (CMT) Acutance was conceived by

Gendron (Gendron, 1973). He defined the CMT as

Definition 2.2.4

$$CMT \text{ acutance} = 125 - 20 \log \left(\frac{200}{MTC_{\text{area(syst)}}}. \right)^2 \tag{2.27}$$

This is a measure of the sharpness that can be produced by the whole optical system, including the observer's eye optics, where MTC_{area} is the area under the MTF curve for a specified component of the system (units mm^{-1}). The MTC_{area} is calculated as follows

$$MTC_{\text{area(syst)}} = \int_0^\infty MTF_c(\mu') \times MTF_f(\mu') \times \cdots \times MTF_0(\mu') d\mu', \tag{2.28}$$

where μ' = spatial frequency and $MTF_c, MTF_f \ldots$ are MTF's referred to the observer's retina. Equation 2.27 was transformed by Overington (1974) to become:

$$CMT \ acutance = 40 \log(6.67 MTC_{\text{area(syst)}}) \tag{2.29}$$

CMT acutance as expressed in equation 2.29 is closely related to MTFA.

MTFA Variants

The Subjective Quality Factor (SQF) was developed by Granger & Cupery (1972). It is based on eye physiology and specifies MTF data in terms of the retinal image, with a bandpass centred on the peak response of the visual system being applied. The function has a high correlation with subjective response. However, Overington asserts that it fails under some combinations of adjacency effects and halation.

Overington (1974) proposes an empirically derived measure of image quality, based on a physical model of the visual system, which he calls visual efficiency. He suggests that using resolution as a quality measure is not appropriate, as its value depends on the conditions of the test environment (*e.g.*, illumination, contrast, shape, *etc*). The definition of visual efficiency is more nebulous than the other MTF-based measures and is defined, in words, as the ratio between the maximum illumination gradient in the retinal image, produced by the combination of the optical and human visual systems and that produced by the "perfect" eye, the maxima being as estimated by the matrix of retinal receptors. The perfect eye has no optical degradation, except for diffraction effects due to diffraction at the pupil and related to its diameter.

In 1987 Barten introduced a new measure, related to the MFTA, which he called the Square Root Integral (SQRI) method. This he defined as

$$J = \frac{1}{\ln 2} \int_0^{\text{Umax}} \sqrt{\frac{M(u)}{M_t(u)}} \frac{du}{u}, \tag{2.30}$$

where u is the angular spatial frequency, Umax is the maximum angular spatial frequency displayed, $M(u)$ is the MTF of the system under test (which is the same as MTF in MFTA).

The function $M_t(u)$ is the modulation threshold of the HVS (which is same as CSF in MFTA). This measure is expressed in terms of just-noticeable-differences (jnds), where 1 jnd is defined as giving a 75% correct response in a two-alternative forced choice experiment (see Chapter 3). Barten argues that the MFTA is very dependent on the visual context. He also points out that the HVS has a linear response to luminance level, whereas a more realistic assumption is that the response of the HVS is non-linear. He claims the SQRI measure largely overcomes these weaknesses.

2.2.5 Entropy Based Measures

The concept of entropy is borrowed from information theory and applied to image measures. Entropy can be considered as the amount of detail or "busy-ness" of an image. Consider a discrete random variable X, with a sample space $X = \{x_i\}$, and x_i occurring with a probability of p_i.

Definition 2.2.5
In a communications channel, the *self information* of any element x_i may be defined for a receiver by

$$I(x_i) = -\log(p_i). \tag{2.31}$$

Definition 2.2.6
Entropy is defined as the average self information by

$$H(X) = -\sum_{i=1}^{N} p_i \log p_i. \tag{2.32}$$

Note,

(i) $H \geq 0$;

(ii) If for some i, $p_i = 1$, then H = 0, and

(iii) Entropy reaches a maximum when all p_i are equal; let i=1 ... a, then the maximum entropy is achieved iff $p_i = 1/a \; \forall \; x_i \in X$. The maximum entropy is then $\max H(X) = \log\{a\}$

If X is continuous then

$$H = -\int_{-\infty}^{\infty} p_x(X) \log p_x(X) dx, \tag{2.33}$$

where $p_x(X)$ is a probability density function and the integral exists.

Independent Pixel Case

Now consider entropy in the context of image analysis. Let $p(x_i)$ be the probability that a pixel assumes an intensity value of x_i in an image X, where there are "a" discrete levels of intensity values, where $x_i \in \{0, \ldots a-1\}$, and we assume that each pixel is independent. Consider an image where all the pixels have the same intensity level, say x_j, then $P(x_j) = 1$. Thus, the image has minimum entropy, which is equal to zero; *i.e.* there is certainty in the value of each pixel, so that the image carries no information. On the other hand, if the occurrence of each the "a" intensity levels is equally likely, then the entropy of the image is at a maximum.

Markovian Related Pixels Case

In general, the pixels in real images are not independent but correlated. To allow for entropy determinations in this context, assume that pixels have a Markovian relationship (Rosenfeld and Kak, 1982). An image can be considered to be a one dimensional (1-D) sequence of pixels as shown in figure 2.3. Here, the probability that a pixel j will have an intensity

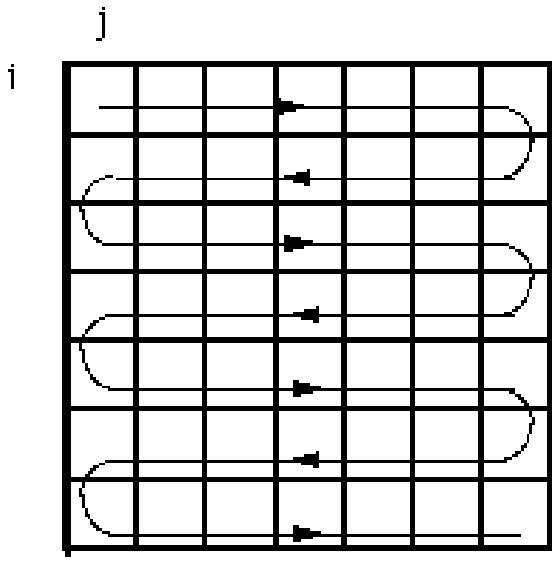


Figure 2.3: An Image Considered as a Markov Chain.

x_j, will depend (under the Markovian assumption) on the intensities of the previous k pixels; *i.e.* $p(x_j|x_{j-1}, \ldots, x_{j-k})$.

Definition 2.2.7
The conditional entropy is now defined (Hamming, 1980) as

$$H(x_j|x_{j-1}, \ldots, x_{j-k}) = -\sum_{X_j} p(x_j|x_{j-1}, \ldots, x_{j-k}) \log\{p(x_j|x_{j-1}, \ldots, x_{j-k})\}, \tag{2.34}$$

where the summation is carried out over all possible realisations of X_j. This gives the uncertainty associated with the intensity of pixel j. Now if the conditional entropy is averaged over all

possible realisations of $(X_{j-1}, \ldots, X_{j-k})$, the total entropy for the Markov chain obtained is

$$H(X) = \sum_{x_{j-1},\ldots,x_{j-k}} p(x_{j-1}, \ldots, x_{j-k}) H(x_j | x_{j-1}, \ldots, x_{j-k}) \tag{2.35}$$

$$= \sum_{X_{j-1},\ldots,X_{j-k}} p(x_{j-1}, \ldots, x_{j-k}) \cdot \left[-\sum_{X_j} p(x_j | x_{j-1}, \ldots, x_{j-k}) \log\{p(x_j | x_{j-1}, \ldots, x_{j-k})\} \right] \tag{2.36}$$

$$= -\sum_{X_{j-1},\ldots,X_{j-k}} \sum_{X_j} p(x_{j-1}, \ldots, x_{j-k}) p(x_j | x_{j-1}, \ldots, x_{j-k}) \cdot \log\{p(x_j | x_{j-1}, \ldots, x_{j-k})\} \tag{2.37}$$

Since (from the definition of conditional probability distributions)

$$p(x_{j-1}, \ldots, x_{j-k}) p(x_j | x_{j-1}, \ldots, x_{j-k}) = p(x_j, x_{j-1} \ldots, x_{j-k})$$

(2.37) can be written as

$$H(X) = -\sum_{X_j,\ldots,X_{j-k}} p(x_j, x_{j-1} \ldots, x_{j-k}) \log\{p(x_j | x_{j-1}, \ldots, x_{j-k})\} \tag{2.38}$$

In computing the entropy of images according to equation 2.38, the higher order statistics of the image(s) must be known up to order k. Here, k determines to what order the entropies are computed. In finding the entropies to order k, it is necessary to determine the k^{th} order grey-level histogram for the k neighbouring pixels. Entropy can be applied as a similarity measure. Consider a reference image X, with entropy H(X), and an object image $\hat{X}$, with entropy H($\hat{X}$). Now the entropy change, given by

$$\Delta H(X, \hat{X}) = H(\hat{X}) - H(X), \tag{2.39}$$

can be used as a measure of similarity.

2.3 Local Image Measures

As already discussed in section 1.3.2 on page 12, in many cases global measurements do not reflect the way in which an interpreter appears to evaluate the quality of an image. The particular elements of the image, upon which the user judges the quality of that image, depend upon the application. (For example, in the case of a cartographic application, edge quality is of paramount importance.) Another important local characteristic of an image is that of texture. Texture analysis can be applied fruitfully to the segmentation and classification in many imaging applications, such as remote sensing.

In the remainder of this section is a summary of the measures used in the literature for firstly, edge quality and secondly, texture analysis.

2.3.1 Edge Quality Measures

It is well known that the HVS is particularly sensitive to the degradation of edge quality, and such degradation has effects on the evaluation of perceived global image quality (Ohtsuka et al., 1988; Ohtsuka and Makoto, 1991).

An edge can be considered as a demarcation line between regions in an image which have different statistical properties associated with their grey-level distributions. Thus, edge quality is important in the context of image segmentation and subsequent interpretation. It has been shown experimentally that errors, produced by low pass filtering of images with high contrast edges, are in fact first detected around the edges in the image (Algazi and Ford, 1980). There are many ways in which the quality of edges can be degraded. Some of these degradations include: edge blur, offset, discontinuities in edge, non- registration of edge points and false registration of edge points.

There has been some effort in the literature to define edge measures, although this has been mainly in the context of image enhancement and the evaluation of edge registration schemes, rather than as quality measures per se. However, these measures can also be applied productively to the task of image quality measurement.

Edge Sharpness

Here mainly non-binary (more than two intensity levels) edge quality is being discussed, as this thesis is addressing image measures for multi-tonal (grey-level) images. Now the characterisation of intensity related edge phenomena will be discussed. The following discussion and development builds upon the work of Panda and Kak (1976) who introduced a measure of edge sharpness which they used to measure the efficacy of edge enhancement in linearly filtered images.

Consider an $n \times n$ image X, which is assumed to consist of a 1-D sequence of pixels as shown in figure 2.3 on page 28. Let the intensity values of each pixel at position $i \in \{1, \ldots, n^2\}$ be denoted by x_i. Now consider a contiguous subset of the 1-D array, call it chain T, such that the first and last pixels of T are from a vertical edge. Now edge sharpness is determined by three factors:

(i) The maximal difference in intensity or *intensity jump*:

$$J(T) = \left| \max_{x_i \in T} x_i - \min_{x_i \in T} x_i \right| \tag{2.40}$$

(ii) The *local mean intensity*, which is the mean intensity of all the edge pixels

$$M(T) = \sum_{x_i \in T} \frac{x_i}{N_T}, \tag{2.41}$$

where N_T is the number of edge pixels in chain T.

(iii) The *boundary width*:

$$W(T) = \varepsilon. \tag{2.42}$$

Consider these factors in relation to figure 2.4. The factors listed are important in determining whether an edge is perceived as sharp or otherwise. It is noted here that there is a distinction between the sharpness of an edge and the thinness of an edge. In the case of the former, the intensity plays a role, while, in the latter case, it is only important to consider whether pixels belong to the edge or not. That is, if the pixels belong to chain T they are edge pixels, and contribute to the edge width if appropriately aligned. In figure 2.4 and item (iii), boundary width corresponds to the property "thinness" just mentioned.

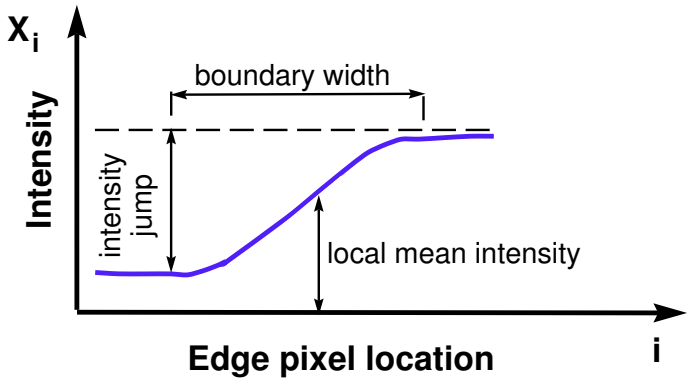


Figure 2.4: Intensity Related Effects and Edge Sharpness.

An intensity jump, as defined in (2.40) and shown in figure 2.4, does not guarantee in itself the presence of an edge. It is always necessary that there exists in a specific direction a sequence of such jumps. This is known as *continuity*, and is an important edge feature.

Taking in to account the factors just discussed in relation to edge sharpness, a reasonable definition of the latter may be expressed as

$$E_{sh} = \frac{J(T)}{M(T) \cdot W(T)} \tag{2.43}$$

The ratio $\frac{J(T)}{M(T)}$ in (2.43) is a measure of edge contrast (Panda and Kak, 1976). Thus the contrast varies as the intensity jump and varies inversely as the mean intensity of the edge. Panda and Kak performed smoothing operations on images and found that in general the intensity jump decreases, while the the width increases and the mean intensity level tends to remain constant. This is reflected in (2.43) by a decrease in edge sharpness. By substituting (2.40) to (2.42), in to (2.43), it is possible to express edge sharpness in these more basic terms as shown in (2.44).

Definition 2.3.1
Edge sharpness may be defined within the region of support of chain T as follows:

$$E_{sh} = N_T \frac{|\max_{x_i \in T} x_i - \min_{x_i \in T} x_i|}{\varepsilon \sum_{x_i \in T} x_i}, \tag{2.44}$$

where x_i is the intensity value of pixel i, ε is the boundary width and N_T is the number of pixels of chain T.

Edge Location Accuracy

Pratt (Pratt, 1978) introduced a measure of edge location accuracy called the "figure-of-merit" which is applied to binary edge maps. This is a relative measure, in that it presupposes a knowledge of a reference edge map. It is also assumed these edges are one pixel wide.

Definition 2.3.2

$$Q = \frac{\sum_{i=1}^{\hat{N}_e} \frac{1}{1+\rho\Delta_i^2}}{\max[\hat{N}_e, N_e]}, \tag{2.45}$$

where $\hat{N}_e$ and N_e are the number of pixels in the test and reference images respectively, ρ is a scaling factor while Δ_i is the distance between edge pixel i in the test image and the corresponding pixel in the reference image. The term in the summation expresses possible shifts of the edge, with the scaling factor ρ providing a relative weighting between blurred edges and sharp or thin but offset edges. If the edge is fragmented or smeared, the value of Q will be lower, since $\hat{N}_e > N_e$, and therefore the denominator term $\max[\hat{N}_e, N_e]$ will be larger. Clearly, $Q = 1$ for a perfect edge.

From the previous paragraph it can be concluded that the "figure-of-merit" gives an overall impression of edge quality by taking into account the various smearing and offset effects. A limitation of this measure is the assumption that the reference edge is only one pixel wide, which will cause difficulties for the assessment of quality in some real images. This problem may be overcome by redefining the "figure-of-merit", with a slight modification, as follows.

Definition 2.3.3

$$Q' = \frac{\sum_{i=1}^{\hat{N}_e^i} \sum_{j=1}^{\hat{N}_e^j} \frac{1}{1+\rho\Delta_{ij}^2}}{\max[\hat{N}_e, N_e]}, \tag{2.46}$$

where ij is the edge pixel location (image co-ordinates) and Δ_{ij} is the distance between the ij^{th} pixel of the test and reference images. In the case of a vertical edge, for example, Δ_{ij} is calculated by relating the first and last pixel of the test image to that of the first and last pixels of the reference image on the i^{th} row. Next the second test edge pixel is related to the second last reference edge pixel, and so on. When the edges are of different widths, the distances for the excess pixels are obtained by relating them to the inter-positions as shown in figure 2.5. By this means, an average distance or midway interpolation between adjacent pixels to the excess ones is obtained.

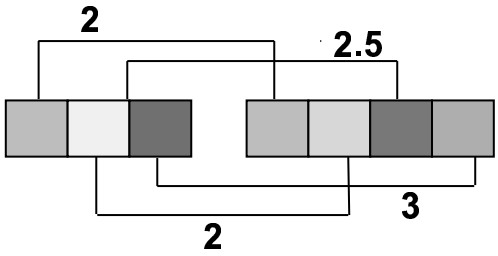


Figure 2.5: Distances for Edges of Different Widths.

Edge Raggedness

In the previous section a measure of the deviation of an edge from its ideal was introduced. However, this measure was not developed to take into account any psychophysical effects that edge dislocations may exhibit to a human observer, who may attach qualities to these effects. One such quality is that of edge raggedness. Hamerly (1981) investigated the effect of edge raggedness on perceived image quality. Assuming the spatial-frequency channel model of the HVS, he found that the phase information was not important and he produced a curve of edge profile threshold resolution for the HVS.

Based on Hamerly's work Gur (1985) developed an objective measure of edge raggedness which was univariate; *i.e.* does not need a reference image. The suggested procedure is as follows. Assume that the image has been binarised and the edge object under consideration (Gur used alphabetical characters) is scanned at right angles to its major axis (see figure 2.6 (a)).

(i) Count the pixels which are internal to the object boundary as a function of scan line to produce the tangential edge profile, $B(n)$ which contains edge and low frequency information (the trend of $B(n)$) about the object;

(ii) determine the low frequency component by cubic spline fitting, to obtain a smooth function $s(n)$;

(iii) find the edge raggedness function $E(n)$ such that

$$E(n) = B(n) - \int_{n-1}^{n} S(x)dx; \tag{2.47}$$

(iv) perform a Fast Fourier Transform (FFT) on $E(n)$ and find amplitudes A_m where $A_m = \sqrt{a_m^2 + b_m^2}$, and a_m and b_m are the amplitudes of the m^{th} harmonics of the sine and cosine (real and imaginary) terms of the FFT respectively;

(v) evaluate the measure of raggedness perception P,

$$P = \int_0^{f_{\max}} A_e(f)df_1 f_1 \sum_m A_{em}, \tag{2.48}$$

where $f_{\max}$ is the upper bound of the HVS's spatial frequency response ($\approx$ 40 cyc/deg), f_1 is the frequency of the first harmonic ($m = 1$) and

$$A_{em} = \begin{cases} A_m - T_m & \text{if } A_m > T_m \\ 0 & \text{otherwise} \end{cases},$$

where T_m is the threshold for the m^{th} harmonic.

P is the shaded area above the T curve in figure 2.6 (b). This perceptual measure of raggedness (P) was not verified by psychophysical studies, but was based upon Hamerly's experimentally derived threshold curve (T) for raggedness perception. Thus it is analogous to the MFTA measure of resolution, which correlates well with human perceptual performance. It is therefore likely that P will correlate well with human perception. However, these studies need to be done.

An assumption that has been made, or ignored, in developing this measure is that the raggedness of the object edges is isotropic. There are probably some processes that produce directionally dependent levels of raggedness. Thus the direction of scanning in the production of $B(n)$ is important. To overcome this limitation, one could determine the $E(n)$'s in various directions and average them, or choose the one with the highest variance. Another factor to consider is the scan rate for digitising the image. It is important that the scanning frequency is above the Nyquist rate for the particular degree of raggedness. For, even if the high frequency components of the raggedness are beyond human perception, the aliasing effects could be perceptible. This requires that some estimate of the high frequency component of the raggedness be made, or - more practically - that some explicit band limiting be applied.

2.3.2 Texture Measures

In considering the development of image texture measures, it is necessary to consider the meaning of texture. Texture conveys information about the spatial distributions of tonal variations within an image region (Haralick et al., 1973). The concept of tone is based on the varying shades of grey of the resolution cells in photographic imagery, which, in a digital image correspond to the intensity[3] levels of the pixels. Texture and tone bear an inextricable relationship to each other and both are present in an image, although one property may dominate the other.

Textures can be characterised by either a structural approach or a statistical approach. A statistical approach is considered in the next section. In a structural analysis, texture can be considered to consist of elements whose shapes, sizes and placement characterise the texture types. These elements or *"primitives"* as they are called are connected regions satisfying certain properties (Wang et al., 1981). Therefore, to describe texture, both the primitives and the placement rules need to be specified.

[3]Intensity here means luminance, in case of active display devices, or illuminance, in the case of hard copy types of reflective image. See Appendix A.

The Grey Level Co-occurrence Matrix

The Grey Level Co-occurrence Matrix (GLCM) plays an important role in statistical texture measures and is thus described here. The GLCM describes the spatial relationships between grey-levels, by calculating the second order statistics of the pixel grey-levels of an image or sub-image. The prominence of second order statistics in current research may be motivated by *Julesz's conjecture*, that the human eye uses such statistics as a discriminator between textures (Gotlieb and Kreyszig, 1990). Consider an image X, of $(n \times n)$ pixels which exhibit a homogeneous texture. Where each pixel x_{ij} has an intensity in a discrete set G, such that $G = \{0, \ldots, a-1\}$, with $a \in \mathcal{J}$ (the set of positive integers).

Consider a pair of neighbouring pixels, each with value $x_{ij} \in G$, where one pixel with respect to the other pixel may have an orientation that is horizontal, vertical or diagonal with respect to the image as a whole. The distance between the pixels can be represented by the distance vector $\vec{d} = [\alpha, \beta]$, with $\alpha = |d| \cos\theta$ and $\beta = |d| \sin\theta$, where θ is the angle between the line joining the pair of pixels with reference to the horizontal line joining the pixels in the row of the pixel under consideration. Clearly, α and β are integer. For example consider the usual case with $|\vec{d}| = 1$. Here the eight neighbours of a pixel starting with the right hand neighbour ($\theta = 0$)and travelling anti-clockwise are designated: [1,0],[1,1],[0,1],[-1,-1],[-1,0],[-1,-1],[0,-1] and [1,-1].

Definition 2.3.4
Let $X_i = \{x_{p,q} : x_{p,q} = i\}$ and $X_j = \{x_{r,s} : x_{r,s} = j\}$; with $|p - r| = \alpha$, $|q - s| = \beta$ and $i, j, p, q, r, s \in G$. Let the number of pixel pairs be denoted by λ. It can be shown that, for horizontal and vertical pixel pairs, $\lambda = m(m - \alpha)$, whereas, for diagonal pairs, $\lambda = (m - \alpha)^2$, for an $m \times m$ image and with $\alpha = \beta$ or 0 and $\beta = \alpha$ or 0. The GLCM $\Lambda[\alpha, \beta]$ is given by

$$\lambda_{i,j}[\alpha, \beta] = |[X_i \cap X_j]|_{\alpha,\beta}\,, \tag{2.49}$$

where $|\ |$ denotes the size of the set.

The matrix Λ is shown in figure 2.7. The distance between the pairs of pixels used in forming the GLCM are limited only by the size of the array. However, in practice, due to computational cost, a distance of one is commonly used. That is, the eight adjacent neighbours of each pixel are used to form the pairs.

Examples

Shown in figure 2.8 are some generalised examples of GLCMs and the associated characterisations of texture images, where $\vec{d} = [1, 0]$. In the case where the GLCM has only one non-zero element on the main diagonal, then all the pixels of the sub-image have the same intensity value. However, if there are more significant values on the main diagonal, say three, then a course texture is implied in the image, with the contrast of the image being increased as a function of the spread of the diagonal elements. If the GLCM is a triangular matrix, say with the elements

above the main diagonal, the pixel intensities in the image will increase from left to right. In the case where the GLCM is symmetric, a "chess board" pattern is implied in the image, with the contrast being a function of the distance of the elements from the main diagonal. A pure "chess board" pattern is only implied when the GLCM has all elements equal to zero excepting two elements, which are necessarily at transpose positions to each other and each is equal to N for an $N \times N$ image.

In summarising, some general characteristics of images can be deduced from their GLCMs.

(i) The more the elements are dispersed with reference to the main diagonal of the GLCM, the finer the texture in the image;

(ii) Contrast in the image is greater when the elements in the GLCM are greater distances from the main diagonal;

(iii) Triangular GLCMs imply images where the intensities of the pixels increase or decrease monotonically in the direction of $\vec{d}[\alpha, \beta]$. Whether the intensities increase or decrease depends upon whether the triangular matrix in question is in the top or bottom, respectively, of the GLCM.

(iv) Symmetrical GLCMs imply "chess board" like structures within the image.

Consider the matrix shown below:

$$X = \begin{bmatrix} 4 & 0 & 4 & 0 & 4 & 1 \\ 0 & 3 & 0 & 4 & 1 & 3 \\ 4 & 1 & 4 & 1 & 3 & 0 \\ 0 & 4 & 0 & 3 & 0 & 3 \end{bmatrix} \tag{2.50}$$

If the top adjacent (12 O'clock) pixel is considered for the computation of the GLCM; *i.e.* $\vec{d} = [\alpha, \beta] = [0, 1]$ then the following matrix is obtained:

$$\Lambda[0, 1] = \begin{bmatrix} 0 & 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 3 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{2.51}$$

From the GLCM various histogram descriptions of texture can be derived. The most direct form is as follows

$$h_{ij}(\alpha, \beta) = \lambda_{ij}/m, \quad i, j \in G, \tag{2.52}$$

where m is the number of pixel pairs.

Texture Measures Based on the GLCM

The GLCM are not usually used directly, but texture features are derived from them by applying certain measures. These measures are designed to reveal some salient characteristics of textures

that correlate with visual percepts. However, Haralick [op. cit.], went beyond this in suggesting 14 textual features ($f_1, \ldots, f_{14}$), some of which have no visual meaning. These 14 features can be put into four broad groups:-

(i) measures that express visual textual properties: second angular momentum (homogeneity) f_1, contrast f_2, correlation (grey tone linear dependencies) f_3;

(ii) measures that are based on statistics: variance f_4, inverse difference moment f_5, sum average f_6, sum variance f_7, difference variance f_{10};

(iii) measures based on information theory (usually entropy): sum entropy f_8, entropy f_9, difference entropy f_{11};

(iv) measures based on information measures of correlation: f_{12}, f_{13}, maximal correlation coefficient f_{14}.

Notation

$P_{i,j}(\alpha,\beta)$ is the i^{th} and j^{th} entry in a normalised GLCM $= \Lambda_{i,j}(\alpha,\beta)/m$, where m is the number of pixel pairs. $P_i(\alpha,\beta)$ is the i^{th} entry in the marginal probability matrix obtained by summing the rows of $P_{i,j}(\alpha,\beta)$; *i.e.*

$$P_i(\alpha,\beta) = \sum_{j=1}^{a} P_{i,j}(\alpha,\beta), \tag{2.53}$$

$$P_j(\alpha,\beta) = \sum_{i=1}^{a} P_{i,j}(\alpha,\beta) \text{ and} \tag{2.54}$$

$$P_k^+(\alpha,\beta) = \sum_{i=1}^{a}\sum_{j=1}^{a} P_{i,j}(\alpha,\beta), \quad i+j=k=2,3\ldots 2a; \tag{2.55}$$

i.e. the relative frequencies of pairs of pixels with intensities i and j, of which the sum of the intensity values is equal to $i+j=k$.

$$P_k^-(\alpha,\beta) = \sum_{i=1}^{a}\sum_{j=1}^{a} P_{i,j}(\alpha,\beta), \quad |i-j|=k=0,1\ldots a-1; \tag{2.56}$$

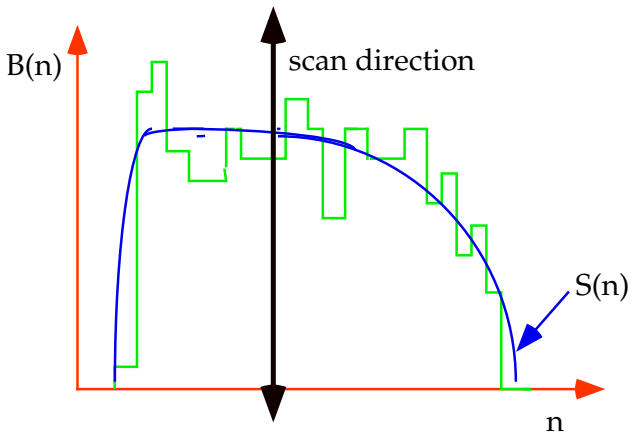
i.e. the set of relative frequencies of pairs of pixels with intensities i and j, such that the absolute difference of the intensity values is equal to $|i-j|=k$, where $\sum_i \equiv \sum_{i=1}^{a}$ and $\sum_j \equiv \sum_{j=1}^{a}$.

Summary of Textural Measures

The following is a summary of the texture measures proposed by Haralick.

(i) Angular Second Moment:

$$f_1 = \sum_i \sum_j P_{ij}(\alpha,\beta)^2; \tag{2.57}$$



(a) S(n) is the low frequency information.

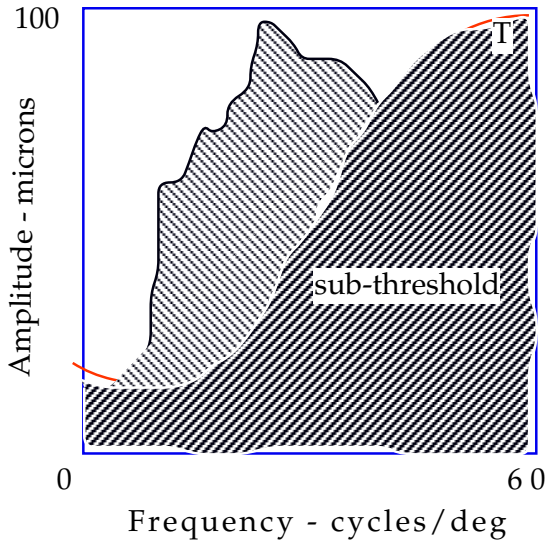


Figure 2.6: (b) Fourier periodogram of $E(n)$, thresholded by the HVS's raggedness resolution curve T.

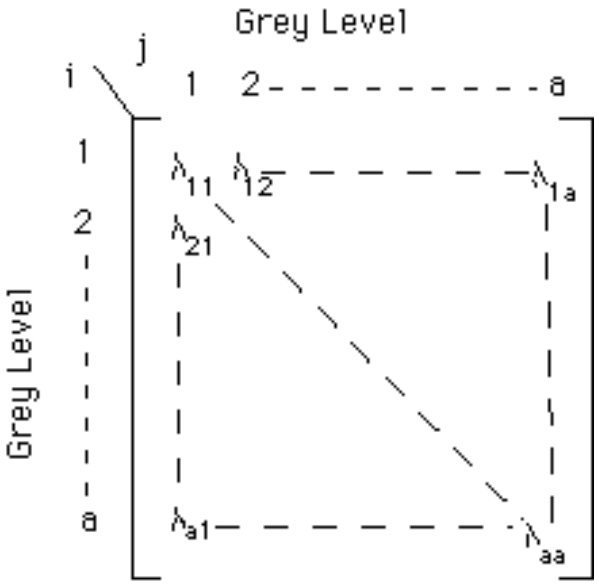


Figure 2.7: The Grey-Level Co-occurrence Matrix.

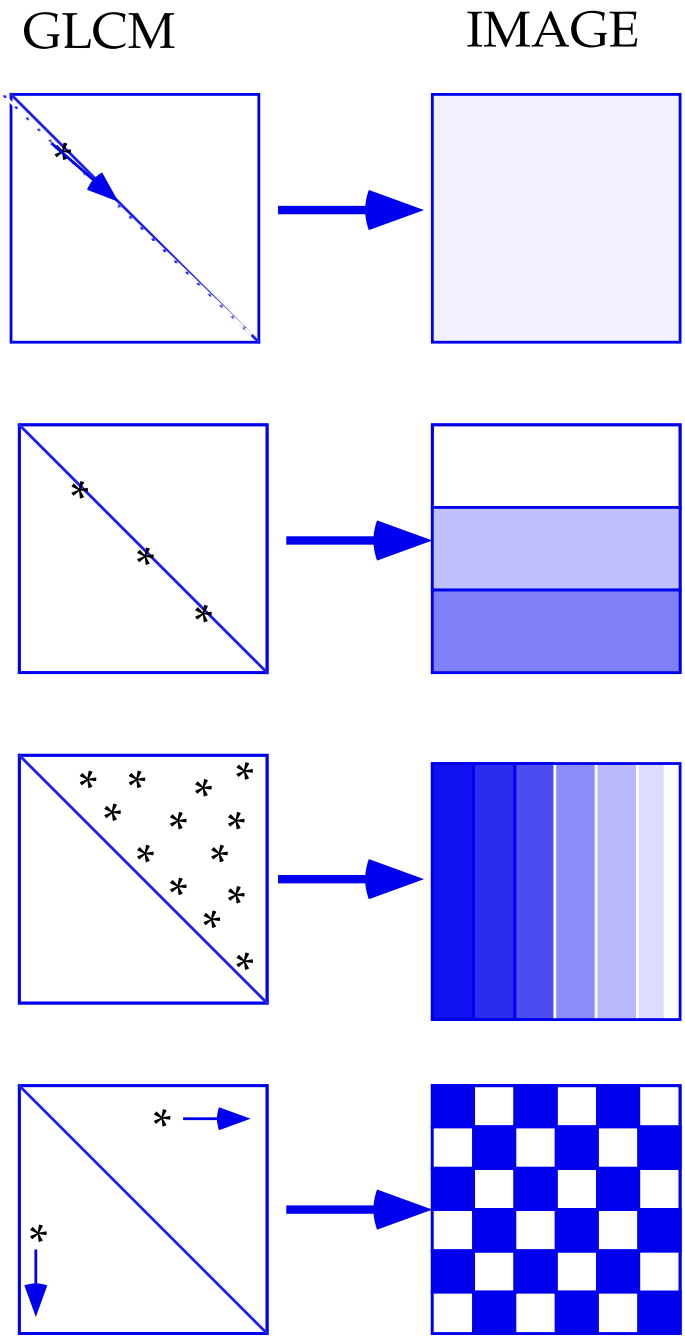


Figure 2.8: Characterisation of Texture from the GLCM ($\mathbf{d} = [1,0]$).

(ii) Contrast:

$$f_2 = \sum_{k=0}^{a-1} k^2 \left[\sum_i \sum_j P_{ij}(\alpha, \beta) \right], \quad |i-j| = k; \tag{2.58}$$

(iii) Correlation:

$$f_3 = \frac{\sum_i \sum_j (ij) P_{ij}(\alpha, \beta) - \mu_x \mu_y}{\sigma_x \sigma_y}, \tag{2.59}$$

where μ_x, μ_y and σ_x, σ_y are the means and standard deviations of $P_i(\alpha, \beta)$ and $P_j(\alpha, \beta)$ respectively;

(iv) Sum of Squares: Variance

$$f_4 = \sum_i \sum_j (i-\mu)^2 P_{ij}(\alpha, \beta); \tag{2.60}$$

(v) Inverse Difference Moment:

$$f_5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} P_{ij}(\alpha, \beta); \tag{2.61}$$

(vi) Sum Mean:

$$f_6 = \sum_{i=2}^{2a} i P_k^+(\alpha, \beta); \tag{2.62}$$

(vii) Sum Variance:

$$f_7 = \sum_{i=2}^{2a} (i - f_6)^2 P_k^+(\alpha, \beta); \tag{2.63}$$

(viii) Sum Entropy:

$$f_8 = -\sum_{i=2}^{2a} P_k^+(\alpha, \beta) \log[P_k^+(\alpha, \beta)]; \tag{2.64}$$

(ix) Entropy:

$$f_9 = \sum_i \sum_j P_{ij}(\alpha, \beta) \log[P_{ij}(\alpha, \beta)]; \tag{2.65}$$

(x) Difference Variance:

$$f_{10} = \sum_{i=2}^{2a} (i - \mu_-)^2 P_k^-(\alpha, \beta); \tag{2.66}$$

where: $\mu_- = \sum_{i=2}^{2a} i P_k^-(\alpha, \beta);$

(xi) Difference Entropy:

$$f_{11} = -\sum_{k=0}^{a-1} P_k^-(\alpha,\beta)\log[P_k^-(\alpha,\beta)]; \tag{2.67}$$

(xii) Information Measures of Correlation:
Let H_i = entropy of $P_i(\alpha,\beta)$ and H_j = entropy of $P_j(\alpha,\beta)$, and

$$H_{ij} = \sum_i \sum_j P_{ij}(\alpha,\beta)\log[P_{ij}(\alpha,\beta)], \tag{2.68}$$

$$H_{ij}^1 = \sum_i \sum_j P_{ij}(\alpha,\beta)\log[P_i(\alpha,\beta)P_j(\alpha,\beta)], \tag{2.69}$$

$$H_{ij}^2 = \sum_i \sum_j P_i(\alpha,\beta)P_j(\alpha,\beta)\log[P_i(\alpha,\beta)P_j(\alpha,\beta)], \tag{2.70}$$

$$f_{12} = \frac{H_{ij} - H_{ij}^1}{\max[H_i, H_j]}, \tag{2.71}$$

$$f_{13} = \sqrt{1 - \exp[-2(H_{ij}^2 - H_{ij})]}; \tag{2.72}$$

(xiii) Maximal Correlation Coefficient:

$$f_{14} = \sqrt{(\text{next to largest eigenvalue of Q})}, \tag{2.73}$$

where

$$Q_{ij} = \sum_k \frac{P_{ik}(\alpha,\beta)P_{jk}(\alpha,\beta)}{P_i(\alpha,\beta)P_k(\alpha,\beta)}.$$

Properties of Texture Measures

It is important to relate the underlying textural processes in an image to the values of the texture measures in order to use them as image quality measures. The following is a brief discussion of some of the properties of the texture measures.

A class of measures, based directly on the GLCM histograms, is that of the mean and variance measures. The variance (f_4) gives information about the contrast within the image or sub-image in which it is applied and the contrast increases with its value. Obviously, if all pixels within the image have the same value, the variance will be zero. Alternatively, the sum (f_7) and difference (f_{10}) variances achieve their minima when all *pairs* of pixels have identical sum and difference intensities, while large sum and difference variances are obtained if there are relatively many pixel pairs with large grey scale sums and differences respectively.

The sum mean (f_6) will have a minimum when the pixels in the pairs in the orientation of $\vec{d}[\alpha.\beta]$ are identical. However, the greater the number of pixel pairs with large intensity sums of the pixels, the larger will be the sum mean. There is an obvious analogous argument for the difference mean.

A closely related measure to the variance and mean is the correlation (f_3). This measure expresses the connection of one value in a pixel pair to the other. There exist in the literature, correlation measures based on information theoretical approaches (f_{12},f_{13},f_{14}), which also express the relation between the pixels, but they are more computationally expensive.

There are measures which express the degree of spread of GLCM values (moments of inertia) around the main diagonal. These include the contrast (f_2) and the inverse difference moment (f_5). The following theorem proved by van der Lubbe, Boxma & Boeker (1984) gives the minimum and maximum values of (f_5).

(i) min $(f_5) = (a^2 - 2a + 2)^{-1}$, where the minimum is achieved iff:

  (a) λ is even: $\lambda_{a1} - \lambda_{1a} = 0$, $\lambda_{ij} = 0$ elsewhere,

  (b) λ is odd: $|\lambda_{a1} - \lambda_{1a}| = 1$, $\lambda_{ij} = 0$ elsewhere.

(ii) max $f_5 = 1$, iff $\lambda_{ij} = 0$, $\forall\,(i-j)$.

Clearly, the minimum of f_5 is achieved only when the absolute difference between each pair of pixel intensities, $|i-j| = a-1$. In the case of the maximum of f_5, the fact that $\lambda_{ij} = 0$, $\forall\,(i-j)$ implies that all pixel pairs are located on the main diagonal of the GLCM.

The minimum and maximum values for the contrast measure f_2 can be derived analogously to that of the inverse difference moment. It is found that:

(i) max $f_2 = (a-1)^2$ and

(ii) min $f_2 = 0$, iff $\lambda_{ij} = 0$, $\forall\,(i-j)$.

Corollary 2.3.1
Transform the pixel intensities such that $i' = xi + c$, with x, $c \in \mathcal{R}$ and $x > 1$, then it holds that

$$f_2^{'} \geq f_2$$

and

$$f_5^{'} \leq f_5.$$

It follows that, as the intensity range is increased, the value of the contrast is increased, while the value of the inverse difference moment is decreased. This is useful as it is consistent with the response of the HVS. The properties of f_2 and f_5 indicate that they are the opposite of each other and it has been shown in the literature that this is generally true. Therefore in most cases, it is enough to apply only one of these measures.

The final group of measures to be discussed here consist of the entropy (f_9) and the angular second moment (f_1). The entropy (see section 2.2.5) measures the "busy-ness" of the texture in the image or equivalently expresses how the GLCM elements are dispersed over the matrix. (Recall that the more dispersed the matrix elements are from the the main diagonal, the finer

[busier] the texture.) In the cases of the sum and difference entropy measures, the diversity of pixel pairs with regards to the intensity sums and differences respectively, are reflected.

While the entropy measures the "diversity" of distributions (entropy is at a maximum when the distribution is the most random), the angular second moment measures the "concentration" of distributions; *i.e.* the homogeneity of the image. In a homogeneous image there are few pixel intensity transitions ($P_{ij}(\alpha, \beta)$ will be concentrated at certain intensities). Hence, the associated GLCM will have fewer entries of large magnitude and f_1 will be relatively large. However, if the image is heterogeneous in intensity values, the GLCM will have a large number of small-valued elements resulting in a relatively smaller value for f_1.

In the literature, there are alternatives to entropy and the angular second moment for measuring diversity and concentration. These have been unified and generalised by van der Lubbe *et al.* (1984), but these will not be discussed here.

2.4 The Application of Image Measures

Generally, image measures can be applied to the evaluation of image processing algorithms, but can also be used as goal functions for image compression, enhancement and restoration techniques. For example, edge quality measures can be applied to measure the performance of edge detection schemes, but may also be used as a goal function from which to develop an optimal edge detector. Hord (1982) mentions how image measures can be used as goal functions in enhancing images for maximal interpretability.

In applying image quality measures for the assessment of image processing algorithms, (*i.e.* the impact of the technique on the image), it is generally advisable to select measures of all aspects of image quality. This will include measures of image similarity (fidelity) and image interpretability (intelligibility), at the global and local levels, with due consideration of the subjective aspects.

The assessment of image quality can be done on a univariate or bivariate (or possibly a multivariate) basis. In the case of the former, the assessment is absolute and based on a single image, while, in the latter situation, a relative measure is obtained from two images, a reference (X) and a test image ($\hat{X}$). In almost all cases, image quality measures are based on some reference, which leads naturally to bivariate measures (*e.g.*, MSE), but univariate measures can be employed if the measurement scale is calibrated by the departure of the test image (measure) from a reference image.

At the moment a growing area of application of image quality criteria is that of image compression. Image compression is becoming increasingly important due to the greater integration of computers and telecommunications (distributed processing), with the growing development of hypermedia, which includes images and their accompanying data storage and processing overheads. In evaluating image (de-)compression techniques many factors come into play, including implementation complexity, real-time processing considerations, compression ratio and the quality of the reconstructed image. Here the concern is the evaluation of image compression

algorithms on a basis of degradation of image quality. In this context the mean square error (MSE) is very commonly used as a measure of quality and MSE is plotted as a function of compression ratio. However, the MSE measures only one aspect of image quality and, as has been shown in this thesis, does not correlate well with human perception. Recently however, researchers have been investigating quality measures that include some of the properties of the HVS (Nill, 1985; Marmolin, 1986; Eggerton and Srinath, 1986; Watson, 1993; Fuhrmann et al., 1995; Avadhanam and Algazi, 1996). Even so, it is necessary to gain the whole 'picture' of image quality by applying a montage of measures which incorporate all (relevant) aspects of quality.

Ultimately the selection of the quality measures depends on the intended application(s) for the image in question and upon consideration of the computational costs of applying the measures.

2.4.1 Global Image Similarity

An impression of the difference between a reference image and a distorted version of it can be obtained by the use of distance measures. An appropriate distance measure to do this is the p-norm which is defined in section 2.2.1. To give a good impression of how similar the distorted image is to the reference image, it is often useful to normalise the distance measure with respect to the reference image, so that the upper value of the measure is bounded to unity.

As mentioned earlier, workers have attempted to obtain a closer agreement between subjective and objective evaluations of image quality by incorporating in the measure (implicitly or explicitly) some form of a model of the HVS. This can result in closer correlation between human and objective assessment of image quality, but one such measure is not consistent over the whole (or even a large part) of the range of image degradations. A disadvantage of this approach is a great increase in computational load; it is likely the extra cost would have to be weighed against the increased agreement with subjective assessment in each particular case. One could apply a normalised cross-correlation measure to gauge similarity. However, due to the complementary properties of distance and closeness, this is not necessary. In implementing such measures, it may be advantageous to use a cross-correlation measure employing the Fast Fourier Transform.

The application of distance measures can be global or piecewise global. That is, such measures can be applied to image blocks, which is appropriate for block image compression schemes. An effective global image measure may be obtained by applying distance measures to sub-regions of the image and combining them in some way. An analysis of the distribution of errors can be facilitated by computing the histogram of the differences between the pixels in the reference and distorted images and by obtaining the values for the errors in individual image blocks, if that approach is used.

Entropy computations give information about the loss of detail in a distorted image. Further, in the context of image compression, entropy measures can be used to determine the maximum possible compression ratio obtainable for lossless compression (via the rate distortion function). As was shown in section 2.2.5, the computation of entropy is done at a large com-

putational cost, due to the higher order probabilities required when using the Markovian image model. However, if entropy is used only as a measure of detail quality, then statistics only up to the second order need to be considered (Caelli and Julesz, 1979).

Edge Quality

The quality of edges is a very important factor in the human evaluation of image quality. Edge quality measures can evaluate the effect of image processing algorithms on the quality of edges. These measures can be based on either binary or non-binary images. In the case of the former, only the pixel position with respect to the edge is evaluated, whereas, in the latter case, the edge pixel intensities are also considered, as discussed in section 2.3.1 on page 30.

All of the edge quality measures discussed in section 2.3.1 emphasise a particular aspect of edge quality. In the practical situation, several of these measures should be applied, but one should be cognisant of the fact that each measure has certain limitations to its application. For example, some edge measures, such as measures of edge sharpness and the 'figure-of-merit' (see 2.3.1 on page 32), are preferentially applied to horizontal and vertical edges. Theoretically, the other edge measures can be applied to edges of any orientation.

All of the edge measures excepting the 'figure-of-merit' are *a priori* univariate measures. However, they can also be used in a relative sense, where the quality of the edges in the reference image X can be assessed and compared with the quality of edges in the test image $\hat{X}$. Application of these measures assumes *a priori* that the true locations of the edges are known and that the edges detected in X and $\hat{X}$ are compatible or structurally similar. This presupposes that the edges in X and $\hat{X}$ have undergone the same edge detection processes; for, if different edge detection algorithms are used (including human edge detection), the results will incorporate information about the performance of the different edge detection procedures, as well as about edge quality.

The two major measures of non-binary edge quality were defined in section 2.3.1; *i.e.* the edge sharpness measure and the local coherence measure. Of these, the former is preferentially applied to vertical or horizontal edges, while the latter can be applied to edges of any orientation.

For binary edges the measures of quality are much simpler than for the non-binary case, making them more attractive to use. The edge properties such as sharpness, mis-location and continuation can be easily measured for both reference and the test images. An exception to this is the 'figure-of-merit', which requires not only the comparison of corresponding edges, but also knowledge about the corresponding edge pixels, making this measure less tractable. However, combinations of other easily applied measures can be substituted for the 'figure-of-merit' to obtain similar information.

Texture Quality

In section 2.3.2, various texture measures were introduced, which were based on the co-occurrence matrix. Because the different measures each address a different aspect of texture quality, it is

necessary in practice to apply various combinations of these measures in order to gain an overall impression of texture quality. For cost effectiveness, it is advisable to take one measure from each class of measures that quantify each particular aspect of texture quality. As examples, one would use either the 'inverse difference moment' or the 'contrast' measure to assess contrast and either the 'diversity' or 'concentration' to measure the homogeneity of the texture *etc.* Many of these measures are directional and should therefore be applied in (say) the four main orientations, to gain an overall idea of the texture quality.

The approach just discussed has been to apply texture measures in an univariate manner, but they can also be applied in a relative sense by comparing regions from a reference image to the corresponding region in a test image. Usually this is done by taking differences between the texture measures obtained for the reference and the test image. Other functions, which incorporate different distance metrics[4], may be used.

2.5 Image Information Measures

Image interpretability, that is the amount of useful information that can be extracted from an image, has already been briefly discussed in section 1.3.1. This is a different aspect of image quality than image similarity (or fidelity), but the two aspects can be related. There are many aspects of image fidelity that correlate with image interpretability. Hord (1982) suggests:-

(i) Minimum noise;

(ii) edge gradient;

(iii) relative structural content;

(iv) fidelity defect;

(v) correlation quality;

(vi) sharpness;

(vii) acutance;

(viii) contrast;

(ix) entropy;

(x) maximum likelihood and

(xi) mean square error.

Image interpretability can be determined objectively, or by subjective testing, and can be applied at either the global or local level, in the same way as similarity measures.

[4] An example of such a distance metric is the Haussdorf measure.

The interpretation of imagery requires the identification and classification of objects and areas within the image under scrutiny. Therefore, the classification accuracy of an image is one measure of image interpretability. This measure is especially useful in the remote sensing context.

2.5.1 Computation of Classifiability Measures

In order to find a classifiability measure, it is necessary to have a reference image, where each pixel for all regions or objects is known. Consider a classifier with input X, which is the

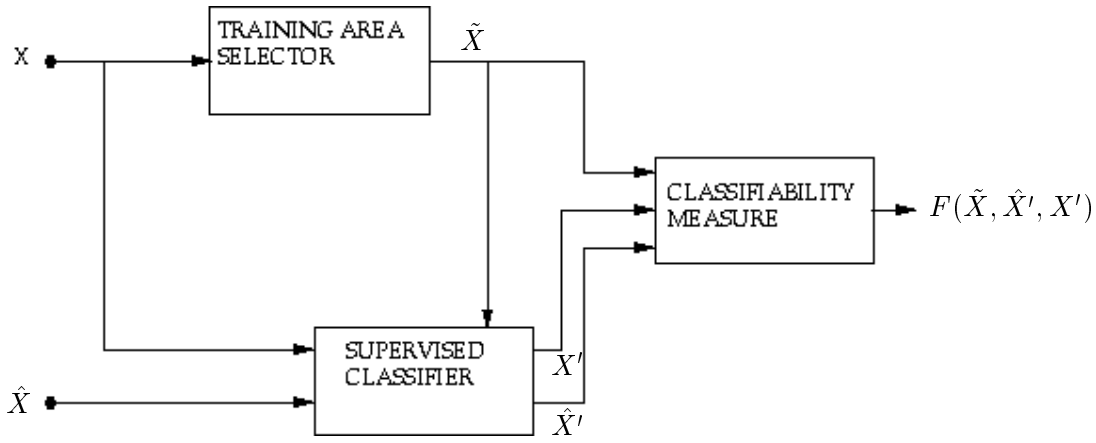


Figure 2.9: Measure of Classifiability

reference image, while input $\hat{X}$ is the test image, for which the quality is to be evaluated. Using this approach, the measure of the quality of $\hat{X}$ is the percentage of correctly classified pixels. However, a "correct" classification is relative to the classified pixels of X, which does not give any relevant information, since these pixels have not been verified. This approach is known in the remote sensing literature as unsupervised classification (Richards, 1986). Further, if the pixels of X have been classified by an interpreter, then the information obtained is relatively biased by the classification technique or algorithm rather than by the quality information contained in X.

These problems can be overcome by applying the method shown in figure 2.9 (after van der Lubbe 1984a), which shows a perfect classifier; *i.e.* one that classifies all pixels of the reference image correctly. Here X is the reference image. A ground-truthed version, $\tilde{X}$, is produced by labelling regions or objects in X. The supervised classifier has as inputs the reference image X, the test image $\hat{X}$ and the labelled image $\tilde{X}$. This classifier has two outputs, X' which is the result of an automatic classification of the reference image X and $\hat{X}'$, the result of the classification of the test image $\hat{X}$.

The image quality (interpretability) of $\hat{X}$ is found by comparing how much information can be extracted from the test image compared to the reference image. In general, the measure of classifiability (interpretability) will be a function of $\tilde{X}$, $\hat{X}'$ and X'; *i.e.* $F(\tilde{X}, \hat{X}', X')$. Classification, in the usual context of remote sensing, relates to pixels found in land use areas.

Here, however, classification is considered in the wider sense, including, for example, detection of edges, textural areas, objects *etc.*

A device often used in evaluating classification is the error or confusion matrix. (See the appendix section B on page 228 for a description.) The confusion matrix gives a description of how well a test image has been classified, compared to a reference image. Different test images can be differentially evaluated against a common reference, provided that the number of classes or objects is identical and the same set of pixels is used.

Classification Accuracy

The device of the confusion matrix can be used as a basis for the definition of classification accuracy measures. It is noted here that the choice and number of pixels for the construction of the confusion matrix is important (see Appendix B on page 228). Let the confusion matrix consist of elements a_{ij}, which represent the number of pixels of class A_j that have been classified as class A_i.

A simple measure of classification accuracy is ratio of the number of correctly classified pixels to the total number of pixels.

$$c_a(\tilde{X}, X') = \frac{\sum_i a_{ii}}{\sum_{i,j} a_{ij}} \tag{2.74}$$

similarly,

$$c_a(\tilde{X}, \hat{X}') = \frac{\sum_i \hat{a_{ii}}}{\sum_{i,j} \hat{a}_{ij}} \tag{2.75}$$

with

$$\sum_{i,j} a_{ij} = \sum_{i,j} \hat{a}_{ij}, \text{ as stated earlier.}$$

From (2.74) and (2.75), a simple measure of classifiability can be defined as:

$$\mathcal{F}(\tilde{X}, \hat{X}', X') = c_a(\tilde{X}, \hat{X}') - c_a(\tilde{X}, X') \tag{2.76}$$

$$= 1 + \frac{\sum_i \hat{a}_{ii} - a_{ii}}{\sum_{i,j} a_{ij}} \tag{2.77}$$

This is a normalised measure and will range in value from 0, where every pixel in the image X is correctly classified while the entire image $\hat{X}$ is incorrectly classified, to 1, where each corresponding pixel in both the images X and $\hat{X}$ are classified identically.

Other related measures have been proposed, such as Jaccard's coefficient (Piper, 1983). This is a special case of (2.77), where the interest is only in pixels of one class, say A_1, being mis-classified as belonging to a second class, say A_2. That is,

$$\mathcal{F}(\tilde{X}, \hat{X}', X') = c_a(\tilde{X}, \hat{X}') - c_a(\tilde{X}, X') \tag{2.78}$$

$$= \frac{\hat{a}_{11}}{\sum_{i=1}^{2}\sum_{j=1}^{2} a_{ij} - \hat{a}_{22}} - \frac{a_{11}}{\sum_{i=1}^{2}\sum_{j=1}^{2} a_{ij} - a_{22}} \tag{2.79}$$

This measure can be demonstrated more easily by reducing the confusion matrix down to a single class problem; *i.e.* a pixel is either classified as belonging or not belonging to a specific class. It has a range from -1 to +1. Shown in table 2.1 is such a matrix, which has been derived from the five class matrix shown in the appendix on page 228. Here class A_1 has been chosen

		Reference classes A_1	A_2	Total	Correct	% Commission
	A_1	35	5	40	88	12
Test	A_2	8	302	310	97	3
image	total	43	307	350		
	% 0mission	19	2			

Table 2.1: A Confusion Matrix for a Single Classes.

as the particular class of interest.

The confusion matrices are produced from a limited set of pixels that have been randomly sampled. Thus, the reliability of the previously mentioned measures depends upon the statistical accuracy of the estimated elements of the confusion matrix. Systematic methods exist to produce confusion matrices with specified reliability (Piper, 1983).

The measures discussed are based on the diagonal elements of the confusion matrix. It seems that not much work has been done with measures that use the whole matrix. More research is needed in this area.

2.5.2 Class Separability Measures

It is arguable that the quality (interpretability) of an image is affected by the statistics of the pixels of the various classes, as these statistics will determine the separability of the classes and hence the interpretability. Methods are therefore required to assess the separability of classes.

Probabilistic Measures

In the current context, each pixel has assigned to it a set of features. These usually consist of the set (vector) of pixel intensity values in some multidimensional space, usually a multi- spectral[5]

[5]Consider a series of sensors which are co-located and 'take' an image of the same area, but each sensor has different spectral sensitivities. The result is a multi-spectral image and each spatial pixel is represented by a vector of intensity values.

space, but can include properties obtained from the neighbourhood (however that is defined) of the pixel under consideration; *e.g.*, textural measures. These feature sets, which have to be analysed in a statistical sense, provide the basis for the differentiation of the various classes or objects. It is reasonable to assume that, the larger the overlap in the underlying feature probability distributions, the higher the chance of mis-classification. Therefore, it is important to introduce rigorous measures of the separability of these distributions, to give some measure of likelihood of mis-classification error. One such measure is known as *divergence* (Richards, 1986). The divergence is defined in terms of the likelihood ratio

$$L_{ij}(\vec{x}) = \frac{p(\vec{x}|\omega_i)}{p(\vec{x}|\omega_j)}, \tag{2.80}$$

where $p(\vec{x}|\omega_i)$ and $p(\vec{x}|\omega_j)$ are the conditional probability distributions for the i^{th} and j^{th} classes respectively. These are shown in figure 2.5.2. Clearly $L_{ij}(\vec{x})$ is a measure of overlap, with $L_{ij}(\vec{x}) = 0$ or ∞ when there is total separation. The divergence of a pair of class distributions

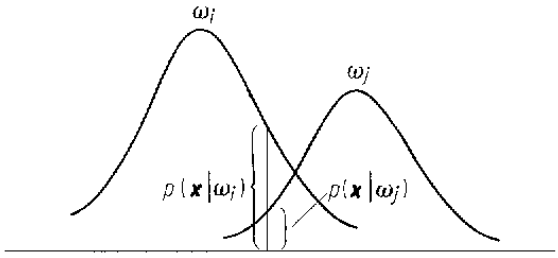


Figure 2.10: Ratio of Probabilities

is defined as

$$d_{ij} =< L'_{ij}(\vec{x}|\omega_i) > + < L'_{ji}(\vec{x}|\omega_i) >, \tag{2.81}$$

where $L'_{ij}(\vec{x}) = \ln p(\vec{x}|\omega_i) - \ln p(\vec{x}|\omega_j)$ and $< \cdot >$ is the expectation operator; *i.e.* $< L'_{ij}(\vec{x}|\omega_i) >= \int_{\vec{x}} L'_{ij}(x) p(\vec{x}|\omega_i) d\vec{x}$, the average value of the likelihood ratio with respect to all feature vectors (possible patterns) in the i^{th} class.

Properties

(i) $d_{ij} > 0$;

(ii) $d_{ij} = d_{ji}$;

(iii) if $p(\vec{x}|\omega_i) = p(\vec{x}|\omega_j)$, $d_{ij} = 0$;

(iv) if $x_1, x_2, \ldots, x_N$ (features) are independent, *i.e.* $p(\vec{x}|\omega_i) = \Pi_{n=1}^{N} p(x|\omega_i) \Rightarrow d_{ij} = \sum_{n=1}^{N} d_{ij}(x_n)$;

(v) $d_{ij}(x_1, \ldots, x_N, x_{N+1}) > d_{ij}(x_1, \ldots, x_N)$.

There is a significant problem with this measure. The probability of correct classification as a function of separation of classes, asymptotes towards one, whereas the divergence increases quadratically. This can give very misleading results.

This problem is overcome with the so called *Jeffries-Matusita Distance* (JMD), also called the *Bhattacharrya Distance*. The JMD between a pair of probability distributions is defined as:

$$J_{ij} = \int_{\vec{x}} \left[\sqrt{p(\vec{x}|\omega_i)} - \sqrt{p(\vec{x}|\omega_j)} \right]^2 d\vec{x} \tag{2.82}$$

which can be interpreted as a measure of the average distance between the class density functions. As a function of separation, this measure shows similar behaviour to that of the probability of correct classification. In the case of normal distributions this function asymptotes towards two (Richards, 1986).

Another separability measure, based on *canonical analysis*, is the ratio of inter-class to the intra-class generalised variances; *viz.*

$$\lambda = \frac{\sigma_b^2}{\sigma_w^2} \tag{2.83}$$

To maximise the separability of classes in feature space, the axes are rotated to produce the

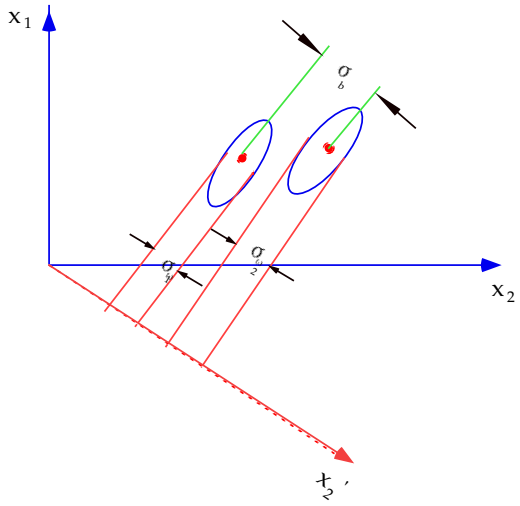


Figure 2.11: Two dimensional example of rotation of axis to maximise separation.

greatest separation of the class means (σ_b^2), when projected on the axis (see figure 2.5.2) and simultaneously showing the minimum spread within the classes ($\sigma_{w_i}^2$).

A major problem with the previously mentioned separability measures is that of their computability, as they require a knowledge of the class conditional density functions. More detailed information concerning the pairwise separability between classes can be obtained by employing distance measures, such as described in section 2.2.1 of this thesis. These metrics

are applied to measure the distance between pixels. The distance between two classes can be computed by averaging the distance between all pixels of one class to all pixels of the other class, or by finding the distance between the class means. The main advantage of these measures is that they are easy to compute.

2.6 Image Clutter Measures

The term "clutter" as used is a general term to describe spatial and, sometimes, spatiotemporal variations in imagery that reduce the availability of target information to a specific sensor. Although other researchers in the area of clutter are interested in man-made electro-optical sensors, this thesis is concerned with the HVS as the sensor.

While the literatures concerned with image quality and clutter measures appear to be largely distinct, it is my assertion that clutter measures are not different to the image measures already discussed in this chapter, except they include a meta-metric to characterise target as distinct from background; *i.e.* the clutter metric has explicitly built into it the concept of target from non-target (background).

The definition given in the first paragraph, implies that the level of clutter has an effect on human visual performance. This is confirmed by numerous studies, including those in this thesis, and particularly the study in Chapter 9, in relation to objective clutter metrics.

2.6.1 Clutter Metrics and Visual Perception

The definition of image clutter explicitly identifies it as a perceptual effect. It is therefore logical to address the problem of deriving clutter measures in the context of models of visual perception. Perception can be defined as the task which governs the transformation of signals into symbols. This signal-in, symbol-out description of the process, is most often considered as a model-based process.

Research in Psychophysics supports viewing perception as a two-stage process (Julesz, 1991). Shown in figure 2.12 is a two-stage model of the visual system from Trivedi and Shirvaikar (1993). The two stages are:

(i) Pre-attentive Processes, and

(ii) Attentive Processes.

The pre-attentive processes take inputs from the sensing mechanisms and generate a set of pre-attentive cues. Examples of these cues are, grey-level discontinuities, texture, depth or motion features. Pre-attention is recognised as primarily a bottom-up process. The processing involved is parallel in nature and does not invoke any models of the scene. Types of tasks which can be considered as typical pre-attentive tasks include feature extraction and figure-ground separation ("segmentation"). The HVS can spontaneously segment regions of an image using a variety of

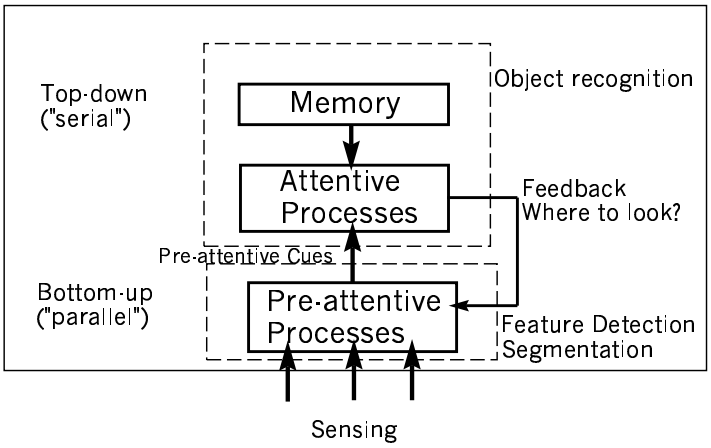


Figure 2.12: A model for visual processing.

image properties, including luminance differences, colour differences, and textural differences; more specifically, differences in the second order statistics (Caelli and Julesz, 1979). This impacts on visual search in the target acquisition process, as it is well established that eye movements during search are guided by scene structure (Yarbus, 1967), and to establish structure requires the achievement of some sort of image segmentation.

The presence of clutter may also induce effects which suppress the visual cues of a potential target. For example, it is well established that there is a suppression of visual sensitivity in the neighbourhood of an edge. Hence edges may be one form of visual clutter. It is also well established that sensitivity to luminance differences depends closely on the level of luminance to which the visual system is adapted (Fechner, 1966; Biberman, 1973; Overington, 1976a; Hall, 1977; Levine, 1985; Falmagne, 1986; Horn, 1986; Buffett, 1986; Moulden et al., 1990) and, further, that the visual system shows strong local adaptation capabilities (Burr et al., 1979). Thus, not only the presence of edges, but their strength, and perhaps their polarity in a local region, could be components of a clutter measure for low level clutter effects.

For the attentive processes, inputs are a set of pre-attentive cues. As opposed to the pre-attentive stage, the attentive stage is classified as a top-down process. It makes use of models of the objects which, it is expected, will be observed. These models are stored in memory and the recognition of objects is carried out by a serial process, in which the memorised models are sequentially compared with objects suggested by the pre-attentive cues. Effects operating through attentive processes are increased confusability of target and non- target shapes, and knowledge-based guidance of search patterns. Objects in a search field are discriminated on the basis of several image properties, of which colour, size, and shape are perhaps the strongest cues (Overington, 1976a). Thus, objects having similar colour, shape and size are highly confusable, whereas those which differ markedly on one or several of these characteristics can be rapidly discriminated. On this view, clutter has a subjective component, in that it requires reference to those characteristics of the target which allow the HVS to discriminate it from the

background. Finally, it should be noted that there exists some "feed-back" from the attentive to the pre- attentive stages. This is important for providing guidance for where to look in the input image.

The above visual system model derives its support from perceptual psychology. It is interesting to note that the model utilised in computer vision also has a very similar layout (Trivedi and Rosenfeld, 1989). A three-level, hierarchical processing framework for model-based vision is generally accepted. Roughly speaking the low-level and intermediate-levels correspond to the pre-attentive stage and the high-level processing corresponds to the attentive stage.

The above models from psychophysics and from computational vision are attempts to describe how signal-level information is eventually transformed into symbolic form by either a human or a machine. There is one very important observation which can be made from these models. There exist at least three different abstraction levels which are encountered in perception. The lowest level is characterised by *signals*, the highest by the *symbols* and in-between are the *pre-attentive cues*. Clutter characterisation studies of the past have primarily focused on the signal level. These studies are influenced by the classical works of Green and Swets (Green and Swets, 1966b), where target detection is modelled as a matched filter design problem. The idea is to use the statistical parameters associated with assumed signal and noise models for developing an "optimum" decision rule for signal detection[6]. The models are based upon the signal-level (lowest) characterisation and are therefore limited with respect to their potential for the quantitative characterisation of perceptual effects.

Instead of the above signal-level solution to the clutter quantification problem, the higher, pre-attentive cue-based solution has been used for characterisation. The specific cue which seems to be quite useful is that of texture. Texture is recognised as an important pre-attentive cue in human and machine perception. It is known that tasks, such as segmentation, are accomplished by using pre-attentive texture-based cues. Indeed many clutter measures are based on texture characterisation, which has already been discussed in section 2.3.2, and some of these measures are discussed later in this section.

We see, then, that the clutter concept is multi-dimensional and to use it effectively for the modelling and prediction of human performance, it must be constrained to apply to those dimensions of prime relevance both to the HVS and to the task addressed. In the remainder of this section, the various classes of clutter metrics and their application are discussed.

2.6.2 Classes of Clutter Metrics

It is not my aim in this thesis to give an exhaustive survey of the literature of clutter metrics, (except in the sense that this has been covered by my survey of image quality metrics), but to give a "flavour" of the types in general use and, in particular, those used in this thesis.

There are many definitions of clutter currently used in the literature of image processing and target acquisition modelling (Schmieder and Weathersby, 1983; Cathcart et al., 1989; Reynolds,

[6]See Chapter 3 for details on signal detection theory.

1990; Doll and Schmieder, 1993; Meitzler et al., 1993; Trivedi and Shirvaiker, 1993; Hilgers et al., 1997), but there is presently no definition known by the author which is clearly the best in all cases or for all images. It is interesting that, while some researchers claim that clutter and noise are different (Schmieder and Weathersby, 1983; Trivedi and Shirvaiker, 1993), others basically equate clutter to noise (Meitzler, 1995; Hilgers et al., 1997). In the case of the former it is asserted "*Noise is recognised as a temporally dependent effect which affects the signal detectability. Clutter, on the other hand, is* [a] *temporally, independent effect.*" (Trivedi and Shirvaiker, 1993). This seems to be making assumptions about the image under consideration. For example, if the image is static, no temporal effects exist, but the image may contain a target embedded in clutter and noise.[7]. Consider an image from a moving sensor, Surely, changes in distance and field-of-view could cause apparent changes in clutter characteristics, due to changes in scale[8]. Schmieder and Weathersby state "... [clutter]*is typically neither stationary, ergodic, nor Gaussian, while noise is usually all of these.*" (Schmieder and Weathersby, 1983). Nevertheless, it is the common experience of image processors that noise does not always obey this edict. There may be times, of course, when the clutter and noise luminance spatial and/or temporal distributions may be quite distinct and well characterised.

As part of the review of the phenomenology associated with clutter metrics, a few of these metrics will be described below. Current commonly used clutter metrics include the following:

(i) Der metric;

(ii) POE metric;

(iii) Schmieder Weathersby metric;

(iv) Texture based clutter.

Der Clutter Metric

Originally the Der metric was devised as a method that could be used to predict the false alarm rate of a given algorithm. The approach taken was as follows: a double window was convolved one pixel at a time over the image. The maximum size of the inner window was set to the expected largest target size. The minimum size of the outer window was set to some value larger than the maximum inner window size. These two features, minimum and maximum, were chosen arbitrarily. At each pixel location, the algorithm decides whether the new pixel is in the same intensity space as the one previously examined and then also whether it fits into the inner window. When an intense region of the image of the approximate target size is found, that region is catalogued. The principle behind the Der method is to multiply the distribution of the target-like areas by the probability of detection distribution. The result should then give the predicted false alarm rate for an algorithm with a given probability of detection distribution. Now, if one simply counts the number of Der objects in the image, that number should indicate

[7] This may have to be ascertained purely from a human subjective viewpoint.
[8] Unless the clutter was of a purely fractal nature, more on this in Chapter 7.

the number of target-like objects in the scene, and, hence, a measure of clutter (O'Kane et al., 1993).

POE Metric

The Probability of Edge (POE) metric is meant to determine the relationship between the human visual detection system and the statistics of the colour or black and white images. First, the image under consideration is processed with Difference-of-Offset-Gaussian (DOOG) filters and is thresholded. This procedure is intended to emulate the early vision part of the HVS (Burt and Adelson, 1983; Peli, 1995). Then the number of edge points is counted and is used as the raw metric. The procedure for calculation is as follows. Firstly, the image is divided into blocks twice the apparent size of the target in each dimension. Then, a DOOG filter as described in (Burt and Adelson, 1983) is applied to each block to emulate one of the channels in pre-attentive vision, with the net effect being to enhance the edges. As discussed in (Rotman et al., 1991), the histogram of the of the processed image is normalised and then a threshold, T, is chosen based on the histogram. The number of points that exceed the threshold in the i^{th} block are computed as $POE_{i,T}$. The POE metric is then computed in a manner similar to the statistical variance technique,

$$POE = \frac{1}{N}\sum_{i=1}^{N} POE_{i,T}^2. \tag{2.84}$$

As Rotman *et al.* (1991) point out, Marr (1982) and other vision researchers have recognised that pre-attentive vision is highly sensitive to edges.

Schmieder and Weathersby (SW) metric

Schmieder & Weathersby (1983) have proposed the concept of a root-mean-square (RMS) clutter metric of the spatial-intensity properties of the background. To date it is the most commonly used clutter measure. The Schmieder and Weathersby (SW) clutter metric is computed by averaging the variance of contiguous square cells over the whole scene:

$$SW = \sqrt{\frac{1}{N}\sum_{i=1} N\sigma^2}, \tag{2.85}$$

where σ^2 is the variance of pixels within the i^{th} cell, and N is the number of cells or blocks into which the image has been divided. Typically N is defined to be twice the length of the largest target dimension. This is a common, though unproved, practice in the clutter metric literature, but is put to the test experimentally in Chapter 7. The signal-to-clutter ratio (SCR) of the image is then given by the average contrast of the target divided by the clutter computed in (2.85).

The variance in (2.85) has been shown by Reynolds (1990) to be equivalent to,

$$\frac{1}{Nk}\sum_{i=1}^{N}\sum_{j=1}^{k}(x_{ij}-\mu)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{1}{Nk}\sum_{i=1}^{N}\sum_{j=1}^{k}(x_{ij}-\mu_i)^2, \tag{2.86}$$

where N is the number of cells, k is the number of pixels per cell, x_{ij} is the radiance of the j^{th} pixel in the i^{th} cell, and μ_i is the i^{th} cell mean radiance. Equation (2.86) was compared by Doll & Weathersby (1993a) with experimental detection times for observers looking at computer generated images of rural scenes with embedded targets. A good correlation between the average detection time and signal-to-clutter ratio (SCR) value was found. One of the fundamental problems of computer-based vision is that the contrast metrics are valid for only a limited group of images.

Texture-based clutter

As mentioned previously, the purpose of defining and quantifying clutter is to aid the development of more realistic human detection models, and texture is an extremely important pre-attentive cue. Clutter is sometimes defined in the sense that areas of similar texture contribute to the distractive capability or clutter of a scene. Texture measures of a scene are potentially very powerful metrics for extracting fine level contrast differences in an image (Meitzler et al., 1993; Doll et al., 1993; Trivedi and Shirvaiker, 1993) and play a crucial role in the modelling of human visual detection.

Texture-based clutter metrics are sometimes called "complexity" metrics in the context of automatic target detection (ATD) and recognition (ATR) systems. The only detailed example of such a metric, to be given here, is the metric by Waldman *et al.* (1988), as this metric was utilised in Chapter 9 with radar derived imagery.

Waldman *et al.* attempt to create an image clutter measure "*... in accord with human intuitive estimates of clutter, being based on the similarity of the background texture to the target in size, shape, and orientation.*". They use the grey-level co-occurrence matrix $Q_{ij}(s, A)$. (This is defined as $\Lambda[\alpha, \beta]$ in section 2.3.2 on page 35, which describes the grey level co-occurrence matrix in detail.) Using their nomenclature, step size is given by s in angular direction A. A probability matrix, P_{ij}, is formed by by dividing each entry of $Q_{ij}(s, A)$ with $Q(s, A)$, "*the total number of possible steps of size s along the direction A.*". That is,

$$P_{ij}(s, A) = \frac{Q_{ij}(s, A)}{Q(s, A)}. \tag{2.87}$$

The sum over all the elements of P_{ij} is 1.

For a constant image with grey-level g, P_{ij} has only one non-zero value on the main diagonal at $i = j - g$. The authors claim that, if an image consists entirely of large texture elements, values of P_{ij} are large for i near j since "*a given step is ... more likely to leave one in the same texture element.*". In fact this is true only if the step size of P_{ij} is smaller than the texture element size and if individual texture elements have a small range of grey levels. They claim that the matrix "*must be decreasing ... for i much different than j.*" This is untrue in many situations. In particular, in the case of large texture elements, there will be an increase in P_{ij} when i and j are grey-levels of adjacent texture elements.

The authors claim that "*some measure of the spread of the co-occurrence matrix about its*

main diagonal could serve as [an indicator of] *the amount of background texture ...*". They suggest the absolute value measure $B(s, A)$ (Pratt, 1978), given by

$$B(s, A) = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} |i - j| P_{ij}(s, A), \tag{2.88}$$

as a good choice. The authors claim that B is large for images with many small texture elements and that B is small for images with a few large texture elements. They point out that B is zero for constant images and for any image when the step size is zero.

The authors claim without proof that $B(s, A)$ increases over s until it equals the average texture size and then flattens out with only small random fluctuations. If this is true, and if texture size is the only image characteristic that will cause this behaviour, then finding the knee on the curve will show the average texture size. The authors claim, again without proof, that this is indeed true. They propose an algorithm for finding the knee on the curve. They define a clutter measure by

$$C_A = (\frac{s}{T_A}) B(s, A), \tag{2.89}$$

where T_A is the known target cross-section at angle A, s is the location of the knee on the curve and B is given by (2.88).

The authors state, without proof, that C_A has a number of properties:

(i) C_A is maximum when the typical texture size is equal to the target size in any direction;

(ii) $C_A = 0$ for a uniform background;

(iii) C_A is invariant under image magnification (scaling).

The second and third properties are undoubtedly true. The first is true whenever $B(s, A)$ increases linearly from zero to a breaking point (the knee) and is then flat. In general, $B(s, A)$ will behave in this manner only in the case of images with a completely uniform, textured background.

The authors tested their hypothesis on an unspecified number of synthetic images, consisting of 60×60 pixels. These images were composed of grids (from 1 to 5 pixels square), with 9 grey levels assigned randomly with a uniform distribution. They show that these images support their hypothesis. They then suggest how to normalise the measurement, given certain theoretical maximum values.

They also claim that their clutter measure is in accord with human intuitive estimates of clutter. They make this claim, based on an application of their measure to a single, highly limited experiment on human perception, using only three subjects.

Finally, they test their metric on a large number (1120) of simple, regular, synthetic images and compare the result to the output of a contrast box ATR. They claim there is a good correlation between the results.

The authors claim that their metric

"... is not just another metric but one of the more fundamental metrics .... We submit this normalised measure of clutter as a very significant clutter metric. It meets all the requirements of a valid metric and should be ranked third in the order of image metrics behind object size and contrast"

Despite their assertions, it appears to this author that Waldman *et al.*, have shown only that their metric is a good measure of clutter when the texture is uniform in size, grey-level, and spatial distribution and there is either a single target or multiple copies of two identical targets that are regularly spaced throughout an image. They did not indicate that they had tested their metric on any real images nor did they compare the result to any ATR in actual use. Images similar to their test cases do not often occur in reality.

Notwithstanding these reservations, this metric was put to the test with real imagery in Chapter 9 and was found to be useful in predicting human visual response in clutter. However, the images used were derived from a radar sensor, which may produce clutter which to some degree matches the description given in the above paragraph.

Chapter 3

Subjective Methods in Visual Psychophysics

Summary: *This chapter provides an overview of psychophysical and experimental psychological methods of analysing subjective visual performance. In particular, the methods used in this thesis are discussed in more detail. Included in the discussion are psychophysical experimental paradigm, rating scale types and methods, analysis-of-variance and signal detection theory. Reference is made to the relevant literature where appropriate.*

3.1 Subjective Methods in the Assessment of Image Quality

Subjective rating by human observers has been commonly used in image quality assessment. The main questions addressed have been: how much does a processed image differ from an original and what is the aesthetic quality impact on the viewer? The latter is important in the entertainment industry, such as television, where the information content of the images is not as important as the overall impression of "niceness" of the picture (Hidaka and Ozawa, 1993). These tests are often performed using "naive" observers. However, trials are often performed using "expert" observers, where the rating of image quality is concerned more with image information and interpretability than aesthetics, such as in military or medical applications.

These approaches to the assessment of image quality or utility are based on an underlying theory of visual psychophysics, which has a long history and a large literature. The rest of the chapter gives a brief overview of applicable psychophysical methods, with some more detail on the procedures used in this thesis.

3.1.1 Psychometric Functions

In any work on assessing human visual performance, the human perceptual response function, known as the psychometric function, is obtained either implicitly or explicitly. The following is a very brief description of this function.

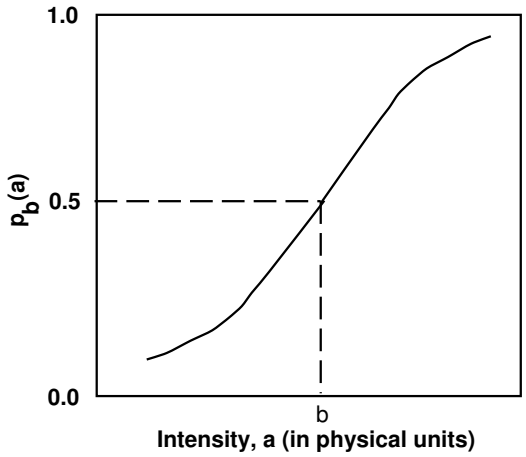


Figure 3.1: Idealised psychometric function.

Consider, for a fixed stimulus b, the probability $p_b(a)$ that stimulus a is judged as exceeding b, where both b and a are in some real interval representing a physical scale (see next section 3.1.3) and where b is a standard of constant magnitude. A somewhat idealised graph of a function $p_b(a)$ which is consistent in its main features with many data, is displayed in figure 3.1. This function is regarded as a function of two variables, where paired values of the stimuli, a and b, map to $p_b(a)$ as shown in (3.3). Such a function $p_b(a)$ is traditionally referred to as a *psychometric function*. This term is also used in a different situation, when $p_b(a)$ may be used to denote the probability of detecting a stimulus a embedded in some "noisy" background b. This is the usual situation encountered in this thesis as described in later chapters. More generally, values of both b and a may vary within the same experimental condition.

A central issue is whether the data support the assumption that two or more psychometric functions are "parallel," that is, can be made to coincide by rigid shifts along the horizontal axis (see Falmagne for a full discussion (Falmagne, 1986)). The rationale for this question is that parallelism is a criterion for an important class of model represented by the equation

$$p_b(a) = F[a - g(b)], \tag{3.1}$$

in which the functions F and g depend on the particular model considered. In other words, any model satisfying this equation must predict parallel psychometric functions. A more general situation can be described by the equation

$$p_b(a) = F[u(a) - g(b)], \tag{3.2}$$

where u is some unknown sensory scale; *i.e.* a probable measure of "sensation". In this case, the psychometric functions are not (necessarily) parallel but may be rendered so by some appropriate transformation u of the physical scale. Obviously, (3.2) generalises to the so called *Fechnerian* equation

$$p_b(a) = F[u(a) - u(b)], \tag{3.3}$$

so named as the form was originally proposed by Fechner in 1860 (Fechner, 1966). The understanding of the issue of parallelism in psychophysical theory is paramount. Parallel psychometric functions indicate that the discrimination (or detection) acuity is uniform on the entire stimulus scale, a fact which may lead to adopting this scale as a measure of sensation magnitude. For instance, in the so-called two-alternative forced-choice design(2AFC, see section 3.1.4 on page 67), the probability $p_b(a)$ is often estimated by averaging the frequencies of the responses in the two alternatives.

The following sections comprise a discussion of the methods used in psychophysics, which are based on an understanding of the issues just discussed.

3.1.2 Experimental Paradigms

There are three main traditional methods used in performing psychophysical experiments. These are:

(i) Method of adjustment;

(ii) Method of limits; and

(iii) Method of constant stimuli.

In (i) the subject manipulates a continuous variable stimulus (*e.g.*, by turning a dial), until it is just noticeable in a detection task. Repeated applications of this procedure yield an empirical distribution of stimulus values, the variability of which is used to compute or estimate the just-noticeable-difference (jnd).

In case (ii) the subject's threshold is elicited by their response to ascending and descending stimulus magnitude sequences, which are controlled by the experimenter. The experimenter varies the value of the stimulus in small ascending or descending steps. At each step the subject reports whether the stimulus appears smaller than, equal to, or larger than the background. The experimenter records the values of the stimulus at which the subject's response shifts from one category to another. This method is used in applied situations, such as audiology, to provide a quick estimate of the point of subjective equality. As pointed out by Levitt (1970), this method has serious defects from the viewpoint of efficiency. The observations may be poorly placed, the estimates may be substantially biased (Falmagne, 1986).

In the method of constant stimuli, case (iii), the stimuli are presented in random or semi-random order, the method is designed to estimate experimentally a number of suitably located points of some psychometric function $p_b(a)$. If a particular mathematical expression is assumed for the psychometric functions (derived, for instance, from a mathematical model), then this expression is fitted to the experimental points. Typically, the mathematical expression of $p_b(a)$ is only specified up to the values of some parameters, which have to be estimated from the data.

If no specific mathematical model is assumed, but the psychometric function appears to be approximately linear, say, between the values .20 and .80, then a straight line can be fitted

to the experimental points in that interval, replacing the mathematical form used above. This has been found to be at least approximately true for the experimental work carried out for this thesis, and, in fact, underlies most of the statistical work that was done. In particular, this is the basis for the *analysis-of-variance* (Peatman, 1964a; Hines and Montgomery, 1980a) (ANOVA) which assumes linear[1] relationships between independent and dependent variables (factors).

Adaptive Methods

Each of the three methods, (i), (ii) and (iii) just described, suffers from one or more of the following defects:

(1) Absence of control on the criterion [(i) and (ii)];

(2) No theoretical justification for important aspects of the procedure [(i) and (ii)];

(3) The estimates may be biased [(i) and (ii)];

(4) Costs; a large amount of data is often wasted (all three methods).

To try and overcome these difficulties, researchers have devised so called adaptive schemes that differ from the methods described previously in that the course of the experiment depends critically on the data: the stimulus presented on trial n depends on the subject's responses on one or more of the preceding trials. At present, none of these methods taken by itself is completely free of defects. However, a suitable combination of methods provides an estimation procedure which seems to be reasonable for empirical applications.

Nevertheless, there is some difference in intent in the use adaptive of methods compared to those methods discussed in sub-section 3.1.2. Adaptive methods are usually employed to find, for a given probability level, a stimulus threshold. Also, adaptive methods often assume some particular shape of the psychometric function. Since this thesis is mainly concerned with the probability of detection versus levels of stimulus properties, adaptive methods were not chosen as an experimental paradigm.

The work carried out in this thesis uses the method of constant stimuli. Some of the reasons for this have already been discussed. Another main reason for using this method is its ability to capture information on the subject's internal criterion when used in a signal detection theory context (see section 3.2.2).

In the following chapters of this thesis, experiments are discussed where observers are asked to rate the quality of images according to some scale. These scaling methods will be basically reviewed in the following subsection.

[1]Of course, the relationships between independent and dependent variable do not have to be linear for ANOVA to work, but its underlying model is linear.

3.1.3 Scaling

The concept of scaling may be defined as the measuring of the psychological response to a physical visual stimulus (image), according to some measurement scale. This definition includes the models, methods and empirical analyses of these response processes. For a full discussion of these issues see Falmagne (1986).

Most measurement involves the assignment of numbers to aspects of objects or events according to some rule or convention. A scale is defined by a group of transformations under which the scale form remains invariant (invariance criterion) (Stevens, 1968). The most commonly used scales include those shown in table 3.1. The foregoing scales represent those in common use,

Scale Type	Allowable Transformations	Examples
Absolute (nominal)	Identity: $X \rightarrow \phi(x) = x$	Numbering of football players.
Ordinal	$X \rightarrow \phi(x) = \psi(x)$, where $\psi(x)$ is increasing monotonic	Hardness scale.
Interval	Similarity: $X \rightarrow \phi(x) = \alpha x$, with $\alpha > 0$	Length, mass
Ratio	Affine: $X \rightarrow \phi(x) = \alpha x + \beta$, with $\alpha > 0$	Temperature
Log-interval	$X \rightarrow \phi(x) = \alpha x^{\beta}$, with $\alpha > 0 \beta > 0$	Density

Table 3.1: Common Scale Types. X is a set of values (which simply may be labels) which are mapped on to a scale via the transformation $\phi(x)$, where x is a particular member of the set; α and β are constants.

with other types possible. Each scale is defined by permissible transformations, that is, those which keep intact the empirical information depicted by the scale. If the scale form has been preserved, the scale form is said to be *invariant*. The following is a basic description of the commonly used scaling methods, but this does not include the methods of analysis. For this see Falmagne (1986) and Stevens (1968).

Absolute Identification.

During a preliminary period the subject is trained to associate a label (say a number $1, \ldots n$) with each stimulus. For the main phase of the experiment, the subject is required on each trial to identify each stimulus as it is presented; *i.e.* produce the appropriate label for each stimulus. Typically, the stimuli are presented randomly and without immediate repetitions.

Category Rating.

This is related to the absolute identification method as each subject is required to assign each stimulus to one of m-ordered categories. These categories are assumed to be have equal separation in a perceptual space; *i.e.* the subjective distance between categories is equal. The number

m of categories is often smaller than the number of stimuli. There exist variations of this method where subjects are instructed to rate pairs of stimuli as to their differences or ratios.

It appears that this method produces very regular data in that the average rating values appear to vary smoothly with an increase in stimulus intensity. This may result in a Fechnerian type relation, which has the form

$$D_{xy} = F[u(x) - u(y)], \tag{3.4}$$

where D_{xy} is the average rating associated with physical intensities x and y, while F and u are monotonically increasing functions.

Magnitude Estimation*.

Magnitude estimation is a very commonly used method of subjective analysis in psychophysics. In the context of subjective image quality assessment it is prevalent in the literature. Here the subject is required to produce "direct" numerical estimates of the magnitude of the sensation caused by the stimulus. According to Falmagne, two variants of the method have been used as follows, with apparently similar results.

(i) The subject is initially presented with a reference stimulus to which the sensory magnitude is assigned some value (modulus), say 100. Other stimuli are then randomly presented, and the subject is given the task of assigning magnitude values such that ratios are preserved; *e.g.*, if a stimulus is perceived to have a magnitude of half the reference stimulus, it is assigned the value of 50. Typically, only a few observations are taken from each subject. These are subsequently combined by computing the geometric mean or median.

(ii) A reference is not provided. The subject is required to assign, to any stimulus presented, a number that seems appropriate as a measure of the stimulus magnitude.

In reading the image quality literature, however, a third variant is evident which is a combination of the elements of methods in items (i) & (ii). In this case, a modulus is given, but no reference, and the subject is required to assign a number, within a predetermined scale, that reflects the quality of the presented image.

Production & Matching Methods.

There is a class of commonly used procedures that come under this heading. In these procedures the subject is required to rate a stimulus by "producing" a value of the sensory variable, by (say) turning a dial. Some examples include the following.

The *magnitude production* method reverses the procedure used for magnitude estimation; that is, the subject is given a number and asked to match this with the appropriate sensory magnitude.

In applying the *ratio production* or *fractionation* method, the subject is required to adjust the magnitude of the stimulus so that it is perceived to be a particular multiple or fraction of a reference stimulus.

The results of applying the methods outlined in section 3.1 can be formalised and summarised in the axioms of the *Krantz-Shepard Theory* which is elucidated by Falmagne (1986).

3.1.4 Methods of Obtaining Subjective Responses

There are three basic methods used in perceptual detection experiments to record observer responses. These are:

(i) The Yes-No procedure;

(ii) The Forced Choice procedure; and

(iii) The Rating procedure.

In a *yes-no* task, the observer is presented with one of two possible, mutually exclusive stimuli per trial. The observer is asked to select one of the two alternatives. There may be either an image containing a target or an image consisting of only "noise". Because the yes-no procedure forces the observer to make a decision based upon a single stimulus in isolation, the observer is in effect making a comparison against an internal model of the stimuli. Therefore, this procedure can be used to probe the observer's internal criterion (bias) when making a decision. The *forced choice* task differs from the yes-no and rating task in that more than one stimulus, usually two (2AFC), is presented within the same interval (temporal or spatial) to the observer, who is required to select only one (target) stimulus per trial . This procedure is useful when the sensory aspects of the task are the focus of interest rather than the observer's criterion.

If one wants to consider the effect of the observer's criterion on decision making, then the yes-no method can be used. However, this is often impractical as it requires many presentations of the stimulus set; at each stimulus set presentation the observers are instructed to adopt a different particular level of confidence. The third method, the *rating method*, can be used to give similar information to the yes-no procedure but only requires a single viewing of the stimulus set. Using this method the observer is required to subjectively rate each stimulus by selecting a value on an ordinal or categorical scale (Metz, 1978; De Ridder and Majoor, 1990).

3.2 Methods of Analysis

Given a set of observer responses taken using a method of the previous section, the next step is to convert these into statistical statements about the observer's discrimination ability. If only the sensory processes are of interest then a forced choice experiment combined with a method of analysis such as ANOVA or multiple regression is appropriate.

Both ANOVA, and a method called *receiver operating characteristic* (ROC) analysis, may be used with yes-no or rating experiments, assuming the experiment is designed to allow for the requirements of both types of analysis (*e.g.*, sections 9.2.3 and 9.2.4 of this thesis). A major difference between the two analysis techniques is that ANOVA focuses on obtaining purely sensory information, which in contrast to the ROC analysis, does not give any information about the observer's internal criterion. In contrast, the ROC analysis retains this information. The next sections give a basic overview of the ANOVA and ROC methods.

3.2.1 Analysis of Variance

A brief description of the ANOVA statistical technique is now given for the sake of completeness. First we need to define a few necessary concepts.

factor An experimental factor is an independent variable, the effect of which we are trying to determine; *e.g.*, background clutter level and target contrast level may be factors in a study.

factor level The factor level is the actual level or value of the particular factor; *e.g.*, clutter level at value "high".

treatment The treatment is a particular combination of factor levels; *e.g.*, clutter level "low" and contrast level "high".

treatment mean The treatment mean is the average value of the dependent variable obtained for a particular treatment; *e.g.*, the mean hit rate obtained over all the subjects for low clutter level with high contrast.

treatment effect The treatment effect, or just "effect", is the difference between a particular treatment mean and the average response score for all the treatments; *e.g.*, the mean hit rate obtained over all the subjects for low clutter level with high contrast minus the mean score over all the treatments.

ANOVA is used in testing the equality of several treatment means. In a typical experiment it would be used to test for any "real" difference in the treatment means for the different treatments. ANOVA does this by dissecting the total variability in the data into its component parts. This variability is due to differences between treatment means for each treatment (*between groups*) and that due to differences in subjective response scores at each particular treatment (*within groups*). The aim of the analysis is to determine whether any of the variability in the data is due to real effects or is just due to "noise"; *i.e.* to error in the data.

Consider the case where the treatment means are in fact equal; *i.e.* we know that the factors in the experiment have no effect on the subjective responses. In statistical terms, it is said that the *null hypothesis*[2] is true. In this case, any variability in the data is due to noise. The within

[2]See Appendix C for more discussion on hypothesis testing.

groups variability ($\hat{e}_{\mathrm{w}}$) and the between groups variability ($\hat{e}_{\mathrm{b}}$) are both estimates of the error. Then the ratio

$$\frac{\hat{e}_{\mathrm{b}}}{\hat{e}_{\mathrm{w}}} \approx 1. \tag{3.5}$$

However, if in fact the treatment means are not equal, the between groups variability will include two components; an error term as before and a component due to the real effect. In statistical terms it is said that the *alternative hypothesis* is true. Then (3.5) becomes

$$\frac{\tau_{\mathrm{e}} + \hat{e}_{\mathrm{b}}}{\hat{e}_{\mathrm{w}}} > 1, \tag{3.6}$$

where τ_{e} is the treatment effect.

The problem is that even though we have found that the ratio defined in (3.6) is greater than unity, how much larger does it have to be to signify a real effect? This question will be answered shortly, but first some more definitions.

The analysis-of-variance is so named since it analyses the sources of variability defined in terms of variance. The sample variance ($\hat{\sigma}$) is defined in general as

$$\hat{\sigma} = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1}, \tag{3.7}$$

where μ is the sample mean and N is the sample size. The variance in ANOVA is also called the *mean square* and is usually defined by:

$$\hat{\sigma} = \frac{\mathsf{SS}}{\mathsf{df}}, \tag{3.8}$$

where SS is the sum of squares and df is the number of degrees of freedom. The numerator and denominator in (3.7) are here equivalent to those in (3.8), but (3.8) is slightly more general. The SS is just the sum of the squares of the differences between two discrete variables or a variable and a constant. Usually the constant is a mean as in (3.7) above. The number of degrees of freedom (df) associated with a variance (or equivalently mean square) corresponds to the number of scores with independent information that enter into the calculation of that variance. As an example, consider the use of a sample mean to estimate a population mean, where the sample is taken from the much larger population. If we want to estimate the population variance as well, we must take into account the fact that we have already used up some of the independent information in estimating the population mean. That is why the df in (3.7) is $N-1$ and not N, as the mean is used in the calculation of the variance and once the mean is defined, only $N-1$ sample values can have any value. Once they are fixed, the remaining value is also fixed.

In the analysis-of-variance, the ratio in (3.6) is usually expressed as:

$$\frac{\mathsf{MS}_{\mathrm{b}}}{\mathsf{MS}_{\mathrm{w}}}, \tag{3.9}$$

where MS_{b} is the between groups mean square and MS_{w} is the within groups mean square. Now we get back to the question, how large does this ratio (3.9) need to be to signify a real effect? This

Summary of all Effects for ANOVA				
Effect	Degrees of Freedom (df)	Mean Square (MS)	F Ratio	p-level
clutter	$a-1=2$	$\frac{\mathsf{SS}_{\text{clut}}}{\mathsf{df}_{\text{clut}}}$	$\frac{\mathsf{MS}_{\text{clut}}}{\mathsf{MS}_{\text{err}}}$	Probability that ratio is due to chance.
contrast	$b-1=3$	$\frac{\mathsf{SS}_{\text{cont}}}{\mathsf{df}_{\text{cont}}}$	$\frac{\mathsf{MS}_{\text{cont}}}{\mathsf{MS}_{\text{err}}}$	
interaction	$(a-1)(b-1)=6$	$\frac{\mathsf{SS}_{\text{err}}}{\mathsf{df}_{\text{err}}}$	$\frac{\mathsf{MS}_{\text{int}}}{\mathsf{MS}_{\text{err}}}$	
error	$ab(N-1)$	$\frac{\mathsf{SS}_{\text{err}}}{\mathsf{df}_{\text{err}}}$	–	–

Table 3.2: Derivation of the ANOVA summary table

problem is solved by referring to the statistical distribution known as the F distribution (Hines and Montgomery, 1980b), of which a theoretical discussion is beyond the scope of this chapter. The ratio in (3.9) is termed the F ratio. In basic terms, the F distribution defines the probability of finding any given F ratio when the treatment means are equal; *i.e.* no real effect. From this, given a specific F ratio we can determine the probability that any apparent effect has arisen purely by chance. Theoretically any F ratio can arise purely by chance, but from the F distribution, we can obtain the confidence to any required level, usually 95%, that a real effect exists; *i.e.* if the F ratio is above the threshold for 95% confidence we say that a significant (real) effect has been found.

Table 3.2 gives an example of how the contents of an ANOVA summary table are quantitatively derived. The exact format of the table depends on the particular design, but table 3.2 gives the derivation for the ANOVA table associated with the design used in Chapter 9, which is the simplest design used in this thesis. In the degrees of freedom (df) column, the three clutter levels are symbolically represented by a, and the four contrast levels by b, with n samples in the overall analysis. The number of degrees of freedom for clutter and contrast are $a-1$ and $b-1$ respectively (and not a and b for reasons discussed in section 3.2.1). For the clutter $\times$ contrast interaction the number of degrees of freedom is the product of the df's associated with each factor involved; *i.e.* $(a-1)(b-1)$. The error term arises from the within-groups variability (3.5) and the number of degrees of freedom is found by pooling the number of degrees of freedom for each group; *i.e.* at each treatment. There are ab treatments each with $N-1$ degrees of freedom each, therefore, the error mean square term has $ab(N-1)$ degrees of freedom.

This technique has many advantages. An important one is that it is well understood and there exist techniques for estimation of the necessary sample sizes (see Appendix C for analysis). There are many good texts available that explain the technique in detail, *e.g.*, (Peatman, 1964a; Hines and Montgomery, 1980a).

3.2.2 Signal Detection Theory and ROC Analysis

The quality of observers' decisions will vary in some manner with their perception of the decision making environment. To be able to measure these effects we need to quantify the types of decisions made by observers as the circumstances change. The approach used for this is divided into two parts:

(i) measurement of the relative frequencies of the observers' decisions against the types of stimuli, and

(ii) based on the pay-offs that result, an evaluation of the benefits that can be gained from the different decision strategies.

Principles of signal detection theory are used to guide the approach, to predict the relationship between the decision performance in various situations and to suggest optimal decision making strategies. In the following section, the basic principles of the ROC will be outlined and referenced by the work of major contributors to the literature (Green and Swets, 1966c; Falmagne, 1986; Goodenough, 1976; Metz et al., 1976; Metz, 1986).

Receiver Operating Characteristics and Methods

Consider a task where an observer is required to determine whether or not a target is present in a noisy image. According to Metz (Metz, 1978; Metz et al., 1976; Metz, 1986), this is done by comparing one's impression of the image with some "confidence threshold" and stating that the target is present only if one's confidence exceeds that threshold. Training and experience would seem to be important factors in determining the proficiency with which the observer can interpret the image data. We will comment later on this issue (section 9.2.6).

Underlying the formulation of the confidence threshold on the part of the observer in a particular situation is a consideration of the costs and benefits associated with each decision. For instance, if the false detection of a target resulted in heavy penalties, then it seems plausible that an observer would consciously or unconsciously raise the confidence threshold. These cost/benefit considerations can be summarised in what is commonly called a *pay-off matrix* (or utility matrix). In the detection task, two types of errors can be made:

Type I the observer may report a target is present when in fact only noise is present (a *false alarm* or *false positive*), or

Type II the observer may fail to report a target that is present (a *miss*).

A correct detection is called a *hit*. The remaining case is a *correct rejection*. The cost/benefit associated with these decisions may be represented as shown in table 3.3.

Let S and N represent positive decisions for the target (signal) and noise respectively. Similarly, s and n denote whether the target is present or absent. Then $p(S|s)$ represents the

Stimulus	Response	
	target S	noise N
present s	hit a	miss b
absent n	false alarm c	correct rejection d

Table 3.3: Cost/benefit (payoff) matrix.

conditional probability of the observer deciding that a target is present when a target is in fact present (a hit) and $p(N|n)$ represents the conditional probability of the observer deciding correctly that only noise is present (a correct rejection). The probability of a false alarm is represented as $p(S|n)$ and the probability of a miss is denoted by $p(N|s)$. The *expected utility* is a measure of the yield of a particular decision strategy, and is defined using the quantities a, b, c and d of the payoff matrix in table 3.3 by (Sperling and Dosher, 1986):

$$U = \alpha_{\mathrm{p}}[ap(S|s) + bp(N|s)] + (1 - \alpha_{\mathrm{p}})[cp(S|n) + dp(N|n)] \tag{3.10}$$

where the true positive fraction α_{p} is the ratio of the actual number of true targets to the total number of observations.

The ROC curve shows the various trade-offs that are possible among the probabilities of the four types of correct and incorrect decisions as the values of a, b, c and d of the pay-off matrix are varied. This in turn induces the observer to adopt different values of the confidence threshold. Specifically the ROC graph is a plot of $p(S|s)$ versus $p(S|n)$. In the military surveillance context, $p(S|s)$ is called *hit rate* (HR) while $p(S|n)$ is called the *false alarm rate* (FAR). An estimate of $p(S|s)$ is the ratio of the number of true positive decisions to the number of actual targets and an estimate of $p(S|n)$ is the ratio of the number of false positive decisions to the number of objects deemed as false targets.

To explain this further, suppose that three pay-off matrices have been used and are denoted by q_1, q_2, q_3 indicating three different confidence thresholds. Let $p(S|s, q_i)$ and $p(S|n, q_i)$, $i = 1, 2, 3$, be respectively the hit and false alarm probabilities. As an example consider the hypothetical data shown in table 3.4.

	FAR $p(S\|n, q_i)$	HR $p(S\|s, q_i)$
q_1	0.10	0.35
q_2	0.40	0.75
q_3	0.60	0.90

Table 3.4: Example ROC data.

The accuracy A of, say, an image analysis procedure, is defined as the ratio of the number

of correct decisions to the number of observations. This is equivalent to

$$A = p(S|s)\alpha_{\mathrm{p}} + (1 - p(S|n))\alpha_{\mathrm{n}}$$

where the true negative fraction α_{n} is given by the ratio of the actual number of false targets to the total number of observations.

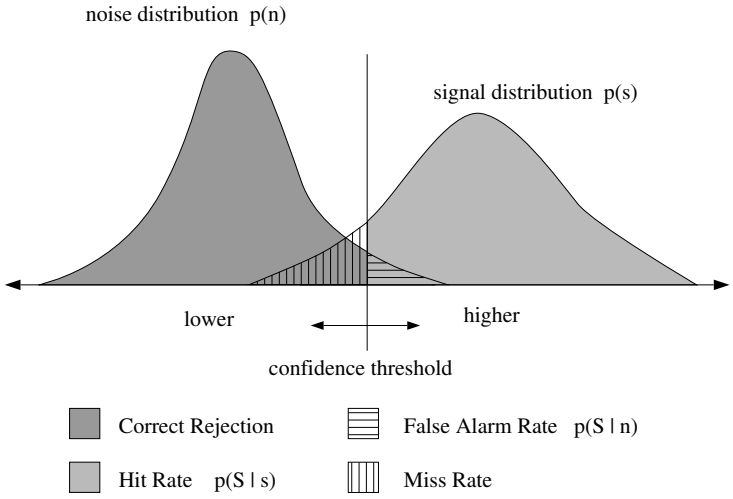


Figure 3.2: A schematic example of the model that underlies ROC analysis. The horizontal axis represents the perceptual response to the quantity upon which decisions are made, while the vertical axis represents probability values. A confidence threshold, represented by the vertical line, separates "positve" decisions from "negative" decisions. The conditional probability of each kind of decision is equal to the area under a distribution on either side of the threshold.

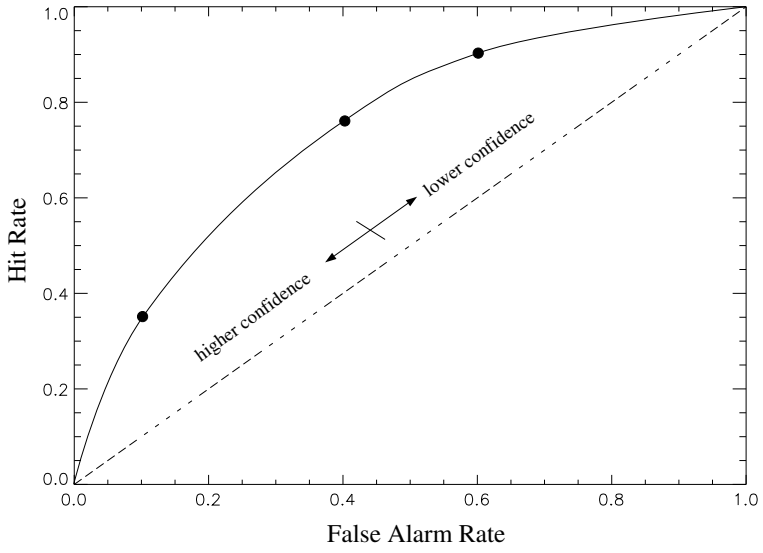


Figure 3.3: The example data with a smooth ROC curve fitted. The dotted line represents a hypothetical ROC curve with no better than chance performance.

These decisions are assumed to be under-pinned by probability distributions as shown in figure 3.2 (Swets, 1973). Strictly, it is not necessary to assume the form of these distributions

in order to carry out an ROC analysis. However, in practice, normality is often assumed to facilitate hypothesis testing and in order to fit smooth curves (using maximum likelihood) to the plotted points on the ROC graph (figure 3.3) (Metz, 1978; Metz, 1986; Green and Swets, 1966a; Burgess, 1989; Hanley and McNeil, 1982; Herrmann et al., 1993). Some useful non-parametric methods have also been used in ROC hypothesis testing and sample size evaluation (Hanley and McNeil, 1982; Rockette et al., 1991). In figure 3.2, $p(s)$ represents the distribution of perceptual responses for image stimuli which contain a target with constant signal-to-noise ratio (SNR), while $p(n)$ represents the distribution for image stimuli which do not contain any true targets (noise only). The threshold, represented by the vertical line, is variable and depends on the observer's estimation of the (prior) probability that the result is positive. This is based on other information such as that derived from information reports and the observer's previous experience.

3.2.3 Calculation of ROC Curve Using the Rating Method

As discussed in section 3.1.4 on page 67, the rating method is often preferred to the yes-no method as it is more economical, yet is conceptually equivalent. In performing an ROC analysis of a visual detection experiment a rating scale such as the following might be used:

Rating	Interpretation
0	Definitely a target is not present.
1	Unsure if a target is present.
2	Maybe a target is present present.
3	Likely a target is present.
4	More than likely a target is present.
5	Definitely a target is present.

This would result in 6 points on the ROC curve which are computed as follows. The point on the curve associated with the strictest decision threshold is calculated by considering only responses at rating 5 as positive and all the rest as negative. For the next point, in the direction of less strictness of the decision threshold, only responses from rating 4 and 5 are considered as positive while the rest are considered negative. This process is continued until any response is considered as positive. This will yield a point at $(1.0, 1.0)$ in the top right hand corner of the graph.

The hit probabilities are computed by dividing the cumulative number of positive responses at each rating level, starting at the most strict, by the total number of stimuli containing targets. Similarly the false alarm probabilities are computed by dividing the cumulative number of negative responses at each rating level by the total number of stimuli without targets. The actual numbers of positive and negative responses at each rating level are determined after the experiment; *e.g.*, the actual number of positive responses at rating 4 will be determined by noting all the stimuli that were given a rating 4 by the observers and determining the total number of stimuli that contained actual targets. An example of the calculations is given in table 3.5.

	Rating						Total
	0	1	2	3	4	5	
Number of hits	3	4	2	2	11	33	55
Cumulative number of hits	55	52	48	46	44	33	55
Number of misses	33	10	6	6	11	2	68
Cumulative number of misses	68	35	25	19	13	2	68
Number of responses at each rating	36	14	8	8	22	35	123
Hit rate	1.0	0.95	0.87	0.84	0.80	0.60	
False Alarm Rate	1.0	0.51	0.38	0.28	0.19	0.03	

Table 3.5: An example of the rating calculations.

3.2.4 Curve Fitting

In order to visualise the ROC curve, fitting to the points can be achieved by eye or by applying an interpolation technique such as cubic splines. However, this is not satisfactory if determination of statistical parameters are required, such as in the case where standard error estimates are required or it is necessary to compute the area under the curve as a measure for comparison of ROC curves.

Usually, the distributions underpinning the ROC analysis are assumed to be normal. In this case a maximum likelihood algorithm can be used to fit a curve to the data points, and this will allow statistical evaluations to be made (Metz, 1986). In effect, this algorithm finds the pair of normal distributions most likely to have produced the given set of data point pairs on the ROC curve (5 in our case).

Similar statistical inferences can be made, including the calculation of the area under the ROC curve, by using non-parametric methods (Hanley and McNeil, 1982). This requires no assumptions as to the underlying distributions for the resulting ROC curve and, in this case, the points on the ROC are joined by straight line segments. If the underlying distributions are Gaussian or near Gaussian, the parametric method is more powerful.

3.2.5 The Area Under the ROC Curve

There have been numerous attempts to produce a single quantitative index from an ROC curve (Hanley and McNeil, 1982), but most of these assume underlying Gaussian distributions as the basis for the observed ROC curve. One of these measures, which is very popular, is the area under the ROC curve, usually denoted by $A(z)$, due to the usual assumption of normally distributed variates. This index varies from 0.5 (no greater than chance) to 1.0 (perfect accuracy) and is usually calculated using an iterative maximum likelihood curve fitting algorithm. This program finds the pair of model parameters, usually expressed as the difference in means and the ratio of variances of the two underlying (assumed Gaussian) distributions, corresponding to the most likely observed ROC curve.

The actual meaning of the area under the ROC curve was first fully addressed by Green &

Swets (1966b), who showed that $A(z)$ was equivalent to the *percentage correct* in a 2-alternative forced choice task. It was pointed out by later authors (Hanley and McNeil, 1982) that $A(z)$ applied equally well to ROC curves obtained by either yes-no or rating methods, and it was shown that $A(z)$ measures the probability that the perceived differences in randomly paired target present and absent images will allow them to be correctly identified. Most importantly, this probability was shown to be equivalent to the Wilcoxon statistic (Peatman, 1964b), which measures the probability that randomly chosen stimuli in target/non-target stimuli pairs will be correctly ranked. This also means that the area under the ROC curve can be estimated non-parametrically; *i.e.* without making assumptions as to the underlying probability distributions. By using this connection, means were developed by Hanley & Mc Neil for assigning error bars to ROC curves (finding the standard error) and for determining required sample sizes. However, they also showed that an estimate for the expected $A(z)$ was required in order to get an estimate of the required sample size that would yield a test of specific statistical power.

Part 2

STUDIES IN HUMAN VISUAL TARGET AQUISITION

Summary: This part of the thesis comprises four chapters detailing experimental work exploring the relationship between image characteristics and human visual performance. It begins by describing an initial experiment which was performed to explore psychophysical approaches to investigating the effectiveness of image metrics for the prediction of subjective responses to image properties. Later experiments are then described, which investigated the effects of controlled degradation on human visual performance in target acquisition and the salient image properties involved. Both static and video images were considered in these experiments. The final chapter, in this part of the thesis, explores the extent to which the size of the clutter area around a target, has a major effect on human visual target detection performance; *i.e.* the "localness" of clutter. This has implications for the application of clutter metrics. In these studies, it was found, in evaluating image quality, that a task-oriented visual performance approach is appropriate, using the probability of detection or response time, rather than the subjective rating of image quality.

Chapter 4

A Study in the Perception of Image Similarity

Summary: *This chapter describes a study which investigates, in a preliminary way, how human observers perceive the same scene information presented either as an infra-red (IR) image or as an optical image. As this was my first visual psychophysical experiment, I also investigated a subjective methodology for producing an interval scale, as a metric of image similarity, and the application of some basic image quality metrics.*

4.1 Introduction

This chapter describes some initial research investigating the application of image metrics to real images. One aim was to explore subjective methodologies, which have the potential to elicit responses from observers, in order to appreciate their perception of image quality. A further aim was to investigate the appropriateness of using common image metrics for real images in a surveillance context, in particular, for measuring the perceived similarity of images. Valuable insights were gained from the study described in this chapter, and these insights assisted in the design of the experiments, which are desribed in later chapters. The results of this study also highlighted some weaknesses in traditional image quality metrics, and showed that a new class of metrics are required in some circumstances.

This study was not conceived solely to meet the research aims just discussed. There was also a driving requirement to satisfy a real world practical need – to be able to cheaply synthesise infra-red (IR) images. Because the production of synthetic imagery based on a full IR physical model is costly and often impractical, an immediate motivation for this study was to gain data that would facilitate the synthesis of IR imagery; *i.e.* produce approximate simulations of IR imagery based on optical imagery of the same scenes. A scene, viewed in both the visible (optical) and IR spectra, contains the same physical objects, but these objects are represented by different distributions of radiant energy (which is visible via an appropriate sensor system) in each spectral range. For example the visual information obtained by shape-from-shading

appears quite different in the visible spectrum to visual information obtained in the thermal IR spectrum. In the case of the former, objects are illuminated by reflective light whereas in the latter case objects are visible (via a sensor) due to emission of IR energy. Nevertheless, the potential information available by both the visible and IR spectrums should be similar. Therefore, it is feasible that IR imagery could be simulated by applying a transformation to the visible imagery; *i.e.* by mapping each IR pixel value to the corresponding visible pixel value. A corollary of this situation, if proven correct, is that humans potentially could obtain similar information from both IR and visible imagery.

In order to obtain an appropriate synthesis of IR imagery, trade-offs have to be made between image "quality" or realism and cost. However, the question arises of how these trade-offs are to be determined. This study, in achieving the stated aims, was intended to make a first step towards answering this question by exploring the efficacy of objective image quality measures for predicting subjective responses to displayed imagery.

In terms of the literature, there appears nothing directly relevant with which to compare this study. The only paper that seems in any way relevant to the study in this chapter is a paper by Toet, Ijspeert, Waxman and Aguilar (1997). Here Toet *et al.* tested whether the fusion of visible and infrared imagery improved observer situational awareness. They found that this was the case. However, there was no attempt to systematically investigate how observers perceive information gained by viewing visible as compared to IR imagery, and thus no direct comparison between this chapter and that paper can be made.

4.1.1 Interval Scale Development by Paired Comparison

This section gives a brief discussion of a method for producing ordinal and interval scales (see section 3.1.3) from a ranking of a set of stimuli, with respect to some quality. This ranking is obtained by presenting pairs of stimuli to a subject and eliciting a response as to which stimulus, in each pair currently presented, has the desired quality.

In the application of this "paired comparison" method, observers are presented with every possible pair of stimuli (images) from the stimulus set and asked to select the one stimulus of each pair that most nearly meets the criterion specified by the experimenter. The procedure is repeated over a sufficient number of observers. No stimulus should be presented more than once within each presented pair, and each stimulus should appear the same number of times in each of the two positions. (In my experiment, three images were presented. However, one was the reference image for comparison with the other two, as shown in figure 4.1.) A matrix is developed from the data, indicating the proportion of times that each stimulus is selected over the other stimulus of the set. Well established assumptions are made about the distribution of the responses of the observers, so that an interval scale can then be developed from the matrix.

The method of data reduction requires the following assumptions:

(i) Observers do not always respond in the same way to each stimulus, but the magnitude of their responses to each stimulus follows a normal probability distribution;

(ii) An observer's response to a particular stimulus is not affected by his response to the alternative stimulus;

(iii) An observer's ability to discriminate the magnitude of the attribute being scaled is equal for each stimulus; *i.e.* stimuli are equally easy to place on the scale.

Example

Ten observers are individually presented with every possible pair of 5 images and asked to select one image from each pair that has the better image quality in some given context. The hypothetical results are tabulated in the form of a proportion matrix as shown in table 4.1. For example, consider the entry "0.30" in row 1 and column 2 which indicates that 30% of the observers rated image 2 to be of better quality than image 1. The column sums indicate the order of the images on an ordinal scale. (Ordinal scales can be more efficiently achieved using rank order methods.)

To develop an interval scale, the proportions must be transformed into Z deviates of the normal distribution as indicated in table 4.2. Under the assumption that that each proportion in the proportion matrix indicates the area under the standardised normal distribution from $-\infty$ to the Z value, the table 4.1 has been transformed to table 4.2. If all assumptions are met,

Image Numbers					
	1	2	3	4	5
1	0.50	0.30	0.20	0.40	0.10
2	0.70	0.50	0.30	0.60	0.20
3	0.80	0.70	0.50	0.80	0.30
4	0.60	0.40	0.20	0.50	0.20
5	0.90	0.80	0.70	0.80	0.50
Col sums	3.50	2.70	1.90	3.10	1.30

Table 4.1: Proportion Matrix. The elements of the matrix are the proportion of times that observers chose the image indicated by the column number over the image indicated by the row number.

the column sums or means of the Z-deviate matrix are the relative positions of the images along an image quality interval scale. Mosteller (1951) developed a method for determining if a scale developed by pair comparison meets the assumptions imposed by the data reduction technique. As the scaling method is based upon changes in the responses of the observers, the differences between the stimuli must be small enough to produce these changes with each pair of stimuli.

Using the above method, the proportions must lie between zero and one. More precisely, a rule of thumb appears to be that the proportions should be between 0.023 and 0.977, so that the Z deviates will range between -2.00 and +2.00. The number of pairs, (N_p), generated by a stimulus set of n stimuli is

$$N_p = \frac{n(n-1)}{2}. \tag{4.1}$$

Image Numbers					
	1	2	3	4	5
1	0.00	-0.53	-0.84	-0.26	-1.28
2	0.53	0.00	-0.53	0.26	-0.84
3	0.84	0.53	0.00	0.84	-0.53
4	0.26	-0.26	-0.84	0.00	-0.84
5	1.28	0.84	0.53	0.84	0.00
Col sums	2.91	5.8	-1.68	1.68	-3.49
Col means	0.58	1.16	-0.34	0.34	-0.70

Table 4.2: Matrix of Z-Deviates from pair comparison data. Elements of the proportion matrix (probabilities) are converted into Z-deviates of the standardised normal distribution.

4.2 Experimental Protocol

A panel of 15 observers was selected opportunistically, all of whom had normal 6-6 vision (with optical corrective devices, if required). The observers had different degrees of experience in judging image quality. The stimuli were presented in such a manner as was outlined earlier under the paired comparison method (section 4.1.1), but in addition, a third reference image was presented with which to compare both images under test. Each viewer sat in front of a Lex 90 display running from a μ Vax computer, and was asked to choose which of two images was more similar to a simultaneously displayed reference image. A typical display is shown in figure 4.1. A series of pairs of images was presented under the control of the observer; *i.e.* the observer indicated his choice by pressing either the right or left arrow key, which in turn initiated the presentation of the next pair of images. The time taken to make each decision was also recorded. Stimuli were viewed with normal room lighting conditions, which were held constant. The viewer was permitted to adjust the viewing distance as he wished.

The rank order data that was obtained from the observers was reduced to obtain an interval scale of relative similarity, following the procedure explained in section 4.1.1. This gave a measure of the relative similarity of the images. It was assumed that the image stimuli were distributed normally along the attribute being scaled. The timing data gave an extra dimension of information which was compared to the ranking data.

4.2.1 The Test Image Set

The images used in this pilot study, were selected from a set of optical and 8-12μ IR images of the same 21 locations. The images were of medium complexity scenes, containing buildings, trees and sometimes cars. The reference image was an IR image of a particular scene, while the pair of images for comparison were both optical images of the same scene, that had undergone some form of intensity transformation.

A subset of 4 of the 21 optical/IR pairs in the original data-set of images was selected. The aim of this selection was to have the image subset contain a cross-section of natural objects and artifacts. This was impossible to fully achieve with only 4 images, but this preliminary

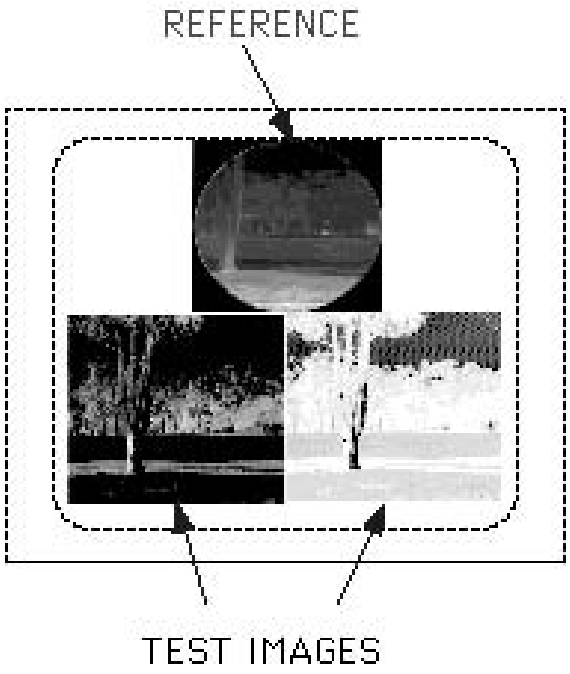


Figure 4.1: Lex90 Display.

study would have soon become unwieldy, with an increase in the subset size, according to equation (4.1). Figure 4.2 shows the four scenes in the subset of 4 images, as obtained through the IR imager. Of this subset, each IR image served as a reference, while each optical image was subjected to an intensity transformation. This yielded 4 images for comparison (including the original optical image). According to equation (4.1), there were 6 pairs for each reference image, and there were 4 reference images, yielding 24 pairs in total for the stimulus set.

4.2.2 Intensity Transformations

The intensity transformations, on the optical images, that were used for this study were:

(i) No transformation (natural);

(ii) Pixel grey-level inversion (inversion);

(iii) Exponential transformation after inversion (exponentiation);

(iv) Logarithmic transformation after inversion.

4.2.3 Application of Image Measures

Four easily implemented measures of image similarity were used. These included:

(i) Mean square error (defined in section 2.2.1 on page 18);

scene 1 scene 2
scene 3 scene 4

Figure 4.2: The four scenes in infrared images.

(ii) E-mean measure (Marmolin, 1986):

$$\text{E mean} = \left[\frac{1}{n}\sum_{i=1}^{n}|M_{x_i} - M_{y_i}|^p\right]^{\frac{1}{p}}, \tag{4.2}$$

where M_{x_i} and M_{y_i} = mean grey-level in a 2×2 sliding window "surrounding" a corresponding pixel in the reference and processed images respectively. This attempted to allow to some degree for the low pass filtering in the HVS (see section 1.2.3 in Chapter 1);

(iii) Difference in entropy (H) where:

$$H = \sum_i p_i \log(\frac{1}{p_i}), \tag{4.3}$$

with p_i, the probability of the pixel occurring. H gives a measure of the amount of information the distribution contains. For a discussion on entropy based measures see section 2.2.5 in Chapter 2.

(iv) Difference in grey-level variance.

These measures were all normalised to yield values between 0.0 and 1.0 .

4.3 Results and Discussion

An ordinal scale of image similarity was obtained by arranging in descending order the sums of the columns of the proportion matrices, which are shown in table 4.3. The indices along the matrices represent the image transformations as shown in table 4.4.

4.3.1 Paired Comparison Rankings

Table 4.5 summarises the rankings of the images obtained by the paired comparison method. As can be seen from the table, the ranking of the images for each scene is different. However, there are some consistencies. For each scene, the images subjected to the inversion-plus-logarithm transformation were ranked the least like the IR image. This is not surprising, since the HVS has a log-like response (see section 1.2.3). Huang (1965) found that image quality was a function of both the number of pixels in an image and the number of quantisation levels for the pixels. For a given size and image resolution then the quality depends on the perceived number of quantisation levels. Since the HVS has a logarithmic response, it will compress further the already logarithmically transformed image, and thus reduce the perceived number of grey-levels even further. However, the other images will be perceived as having relatively more quantisation levels and thus be seen as more similar.

Surprisingly, based on the rankings shown in table 4.5 on the next page, in scene 1 and scene 2 the natural images were considered closest to the IR images, while, in scene 3, the natural image was ranked second only to the inverted image. Scrutinising the four scenes shows that

Image Numbers								
	1	2	3	4	1	2	3	4
	Scene 1				Scene 2			
1	0.50	0.60	0.13	0.70	0.50	0.33	0.13	0.77
2	0.40	0.50	0.20	0.70	0.67	0.50	0.27	0.67
3	0.87	0.80	0.50	0.83	0.87	0.73	0.50	0.80
4	0.30	0.30	0.17	0.50	0.23	0.33	0.20	0.50
Col sums	2.07	2.20	1.00	2.73	2.27	1.90	1.10	2.73
	Scene 3				Scene 4			
1	0.50	0.27	0.27	0.57	0.50	0.57	0.10	0.23
2	0.73	0.50	0.50	0.53	0.43	0.50	0.10	0.20
3	0.73	0.50	0.50	0.63	0.90	0.90	0.50	0.80
4	0.43	0.47	0.37	0.50	0.77	0.80	0.20	0.50
Col sums	2.40	1.73	1.63	2.23	2.60	2.77	0.90	1.73

Table 4.3: Proportion Matrices. Indices: 1,2,3,4 represent the image transformations. 1: grey-level inversion (inv), 2: inv + exponentiation, 3: inv + logarithm, 4: natural

1:	grey level inversion;
2:	inversion & exponentiation;
3:	inversion & logarithm;
4:	natural.

Table 4.4: Table defining the code (indices) used to represent the image transformations.

scene1	scene 2	scene 3	scene 4
4213	4123	1423	2143

Table 4.5: Subjective rankings of perceived similarity between a transformed visible image and the reference IR image, which were obtained from the pair comparison data. Indices: 1,2,3,4 represent the image transformations. 1: grey-level inversion (inv), 2: inv + exponentiation, 3: inv + logarithm, 4: natural

MEASURE	scene1	scene 2	scene 3	scene 4
MSE	2143	1234	2143	2413
EMEAN	2143	1234	2143	2413
ENTROPY	2341	1324	2341	3241
VAR	2341	1432	3214	2341

Table 4.6: Objective rankings obtained from the values "measured" by the image metrics of the similarity between a transformed visible image and the reference IR image. Indices: 1,2,3,4 represent the image transformations. 1: grey-level inversion (inv), 2: inv + exponentiation, 3: inv + logarithm, 4: natural

they all have different contents. In addition, scenes 1 and 2 are heavily shadowed, scene 3 less so and scene 4 not at all. It seems likely that shadows had a marked effect on the perceived similarity of the images for each scene. This may be due to the fact that, when IR images contain sun and shadow, the intensity gradients in these areas are in the same direction as in natural visible images, whereas the inverted image intensity gradients in these areas are in the opposite direction.

The results for the objective image measures are tabulated in table 4.6 on the preceding page. None of these rankings seemed to correspond with the subjective rankings. This indicates that the HVS does not use global statistics as the major means of deciding on image similarity.

4.3.2 Latency and Pair Similarity

Table 4.7 shows the response time for each set of image pairs that correspond to each reference image. Based on the hypothesis that the latency will increase with increased similarity of the two test images presented, these data provide information on the degree of perceived similarity of images with different intensity transformations. Under the hypothesis just mentioned, a ranking of perceived similarity of image pairs was set up accordingly for each image set; *i.e.* the 6 pairs of stimuli obtained per reference image according to equation (4.1) on page 80. Each of the 6 pairs of numbers in a set represents a pair of images that were compared for perceived similarity. The actual numbers in each pair are codes which represent the image transformations according to table 4.4 on the page before.

image set (for scene) 1: 1,2; 2,4; 2,3; 3,4; 1,4; 1,3.
image set (for scene) 2: 1,4; 1,2; 2,4; 3,4; 2,3; 1,3.
image set (for scene) 3: 1,2; 1,4; 1,3; 2,4; 2,3; 3,4.
image set (for scene) 4: 1,2; 2,3; 1,4; 1,3; 3,4; 2,4.

Image Numbers								
	1	2	3	4	1	2	3	4
	Scene 1				Scene 2			
1	0.00				0.00			
2	11.49	0.00			5.64	0.00		
3	6.33	7.56	0.00		3.34	3.58	0.00	
4	6.95	10.99	7.36	0.00	6.62	5.33	4.18	0.00
	Scene 3				Scene 4			
1	0.00				0.00			
2	6.94	0.00			6.28	0.00		
3	5.92	5.22	0.00		4.05	5.45	0.00	
4	6.59	5.65	5.07	0.00	5.12	3.86	3.95	0.00

Table 4.7: Pair Latencies in seconds. Indices: 1,2,3,4 represent the image transformations. 1: grey-level inversion (inv), 2: inv + exponentiation, 3: inv + logarithm, 4: natural

This ranking of similarity between image pairs was also obtained from the z-deviate matrices as shown in table 4.8. The sums of the columns of these matrices represent the values on an interval scale. Thus similarity between pairs was obtained by subtraction. The rankings thus

obtained were as follows:

image set (for scene) 1: 1,2; 2,4; 1,4; 1,3; 2,3; 3,4.
image set (for scene) 2: 1,2; 1,4; 3,4; 2,4; 1,3; 2,3.
image set (for scene) 3: 2,3; 1,4; 2,4; 3,4; 1,2; 1,3.
image set (for scene) 4: 1,2; 1,4; 2,3; 2,4; 1,3; 3,4.

Image Numbers								
	1	2	3	4	1	2	3	4
	Scene 1				Scene 2			
1	0.0	0.26	-1.12	0.53	0.0	0.18	-1.28	-0.74
2	-0.26	0.0	-0.84	0.53	-0.18	0.0	-1.28	-0.84
3	1.12	0.84	0.0	0.96	1.28	1.28	0.0	0.84
4	-0.53	-0.53	-0.96	0.0	0.74	0.84	-0.84	0.0
Col sums	0.33	0.57	-2.92	2.02	1.84	2.30	-3.40	-0.74
	Scene 3				Scene 4			
1	0.0	-0.61	-0.61	0.18	0.0	-0.44	-1.12	0.74
2	0.61	0.0	0.0	0.08	0.44	0.0	-0.61	0.44
3	0.61	0.0	0.0	0.33	1.12	0.61	0.0	0.84
4	-0.18	-0.08	-0.33	0.0	-0.74	-0.44	-0.84	0.0
Col sums	1.04	-0.69	-0.94	0.59	0.82	-0.27	-2.57	2.02

Table 4.8: Z-deviate Matrices. Indices: 1,2,3,4 represent the image transformations. 1: grey-level inversion (inv), 2: inv + exponentiation, 3: inv + logarithm, 4: natural

These rankings agreed reasonably well for the first few pairs with the timing data. This gave some support for the hypothesis that the timing gives information on the similarity/confusion between images, but is probably not a good measure overall of image similarity in this context. The task employed here involved a reasonably high level of cognitive involvement, in that the visual task required interpretation on the part of the subject. That is, they had to make judgement about image similarity, which requires cognitive input rather than just low level visual processing. Because of the high level of task complexity, it may be that response time is not an appropriate measure for this type of task. As will be shown in later chapters, using lower level (early vision) detection tasks, response time turns out to be generally a very accurate measure of performance, but with some caveats (which are discussed in Chapter 5).

4.4 Conclusions and Implications for Further Work

From the results of this experiment it is apparent that the content of the image scenes is important in how the image is perceived. It has been shown that relationships of the regions within the scene can be very important, as was illustrated with the regions of shadow and sunlight. These relationships could not be measured by means of global[1] statistical measures, as shown by the results of this study. That is, the common global image metrics of image similarity that were

[1]See Chapter 1 section 1.3.2 on page 12.

used in this study, did not correspond to human subjective perceptions of image similarity. This strongly suggests that these metrics are inappropriate for use in the surveillance context, since the stimuli that were used here are real world images that are typical of surveillance imagery.

To capture the complexity of the images, measures of local (region based) image properties are required and to "measure" the relationships between image objects, syntactic[2] or semantic (Gonzalez and Wintz, 1987a) type of measures may be useful. The latter type of measures are beyond the scope of this thesis. However, a proposed system for image quality assessment, which uses this type of measure, is discussed in the final chapter, (Chapter 10) in the context of further work.

Because of the reasons outlined in the foregoing discussion, the measures applied here are not able to address the following complexities possible in the synthesis of IR imagery from visual imagery.

(i) Incorrect grey-level of regions or objects;

(ii) Unexpected irregular regions of high or low intensity, due, for example to areas unexpectedly covered in water or shadows;

(iii) Incorrect shape and type of natural vegetation;

(iv) Incorrect shape of artificial structures; *e.g.*, rectangular prism generic shape for building of different shape or which has added structures.

Therefore it is highly unlikely that a mapping can be made from pixel values of a visible image to the equivalent IR image, except in a contrived situation. The consequence of this statement is that the synthesis of IR imagery will require physical modelling; *i.e.* it is unlikely that a tractable algorithm exists to convert visible images to synthesised IR images of the same scene.

The results obtained in this study were not definitive. However, the study gave me some valuable insights into the use of appropriate subjective methodologies for visual psychophysical experiments. Keeping in mind that this study was only preliminary, the limitations may be to be partly due to the complex nature of the visual task, and partly due to the constraints of the experiment itself. In the case of the former, the subjects were required make a judgement (rating) about image similarity, which requires a high level of cognitive input, while in the latter case the experimental complexity did not match the subjective task complexity. That is, only 4 scenes with only 4 image transformations were used, and only simple global image metrics were used for comparison with the subjective responses. Each of the 15 subjects viewed each stimulus (treatment) only once. This did not allow for inter-subject variability, which may have been large for such a complex task.

Because of the reasons just given, this study convinced me that further experiments should be based on well-defined visual tasks, such as detection or recognition, in order to determine directly the effects of image properties on performance; *i.e.* to determine the image quality in

[2]That is structural and relational aspects of the objects within the image scene itself.

terms of utility rather than nebulous qualities such as aesthetics (see Chapter 1 section 1.1.1 and Chapter 3 section 3.1 which discusses the issue of image utility versus image aesthetics). The use of experiments based on visual task performance also facilitates the design of experiments to allow for both within-subject and between-subject variablity. This is the case when employing visual task based experiments, such as target detection, because the time to present each stimuli and gain a response is short, which allows the presentation of many stimuli, including repetitions, to each subject.

Chapter 5

The Effects of Image Compression on Human Target Detection

Summary:

The aims of this chapter include analysing human visual performance under a well defined visual task, with imagery that has undergone degradation. Image compression was chosen as the means of degradation because it is precisely controllable and image compression has practical application.

Two methods of static image compression - JPEG and a fractal-based method were compared in terms of the detectability of simple targets following compression and decompression of the images containing such targets. Targets consisted of rectangles of various sizes and contrasts, which were embedded in images of natural terrain. Using compression ratios of from zero to thirty five, it was found that the loss in detectability of targets in images compressed using the fractal technique was significantly greater than the loss for the JPEG-compressed images.

5.1 Introduction

In Chapter 1 it was stated that image quality can be viewed from a purely aesthetic point of view or it can be analysed in the context of specific visual tasks. In the latter case, this has implications for the image measures used and the method of subjective evaluation. This chapter considers the analysis of human visual performance under a well defined visual task, with imagery that has undergone degradation. It explores methods for the evaluation of human visual performance, including probability of detection, response time and subjective rating. Image compression was chosen as the means of degradation because it can be controlled precisely and has practical application arising from the current explosion in digital imagery.

Image compression is becoming increasingly important due to the greater integration of computers and telecommunications, with their increasing demands on digital storage and transmission systems. In many cases, the overriding concern is the quality of the reconstructed image.

Therefore, there is a need to answer the question of how much compression can be achieved (what is the minimum image quality required) to achieve a certain task, in the context of constraints upon image storage and/or transmission. Consequently, when considering an imaging system for a well defined visual task, such as target detection, it is important to first assess to what extent the compression module is likely to affect user performance on the given tasks. If possible, compression schemes ought to be tuned to the specific task(s) which will be carried out with the images in question.

Lossless image compression can be attained, but this achieves relatively low compression ratios (cr); *i.e.* about 2 - 5 :1. If higher cr are needed, then lossy compression is required. The human visual response is sometimes considered in lossy compression, in that an emphasis is put on removing the information that is less visible to the human observer. However, this usually still results in reduced image quality, especially at high compression ratios. (See section 5.1.1 for a discussion of compression methods.) The effect of this loss in image quality will depend upon the task for which the image is used.

There are several well known methods for compression of still image frames - including the JPEG standard (Wallace, 1992; Leger et al., 1991), which is based on partial representation of the Fourier coefficients in local regions of the image, and the method developed by Barnsley, based on the theory of iterated functions ("fractal" technique) (Fisher, 1993; Jaquin, 1993; Lu, 1993). The evaluation of much image compression, in terms of image quality, has been based on subjective evaluation techniques (Van Dijk and Martens, 1997; Kaukoranta and Nevalainen, 1996; Fuhrmann et al., 1995; De Ridder and Majoor, 1990), with criteria attuned more to the psychological constructs of Gestalt theory, rather than to the suitability for specific tasks.

There are good reasons to believe that the relative effectiveness of various compression schemes would be task dependent. For instance, if a task involves the detection of only small features in an image, then it may not be necessary to retain the low frequency components in a compression scheme, since typical point target detection algorithms for automatic target detection eliminate all low frequency components.

A common surveillance requirement is the visual acquisition of human targets in infrared imagery of natural terrain. This is particularly a requirement for the Military, Customs, and Law Enforcement agencies, and there is a growing realisation that imagery from such surveillance may need to be compressed for transmission and storage. It is therefore appropriate that compression procedures be examined in terms of the intended use of the data being compressed. This chapter describes psychophysical experiments designed to assess the degree to which two different compression schemes affect such a task.

5.1.1 Compression Methods

There exist two basic schemes to perform compression. The first general approach is to exploit redundancy in the data in its original domain, while the second approach transforms the data to a new domain (space) which reduces the correlation in the data.

In the first method, statistical redundancy, which is associated with the amount of predictability in the data, is removed. For example, if it is known that an image consists of pixels, all of the same grey level, then every pixel after the first contains redundant information; that is, the information in every pixel is predictable, or the image has a low *entropy*[1]. If, on the other hand, the image consisted of pixels of totally random grey level, the information gained from any one pixel will not give any information about any other; in other words, the entropy will be high (at a maximum for a random field). This approach often leads to *lossless compression*, where all information is retained, as the result of such methods as entropy or run length coding. Unfortunately, lossless methods do not achieve very high compression ratios.

The second general approach is to transform the original data in such a way that a maximum amount of information is compacted into a minimum number of samples. Usually, after transformation, the number of bits of information, which are allocated to each datum sample, is related to how much of the original information (energy) is contained in that sample; *i.e.* the samples are quantised. It is the quantisation step that allows compression, with the transformation step being lossless and thus completely reversible. The result of this is *lossy compression*, since some information is lost. However, often the selection of the information to discard is performed with the response of the HVS in mind, so as the minimise its impact for human observers. In this regard, usually only the contrast sensitivity function of the HVS is considered. This is the case in JPEG compression for example (Wallace, 1992). With lossy compression, very high compression ratios can be achieved.

There are many schemes for image and video compression which take either or both of the two general approaches just discussed. Figure 5.1 displays a hierarchical diagram of the main types of compression schemes in use. Some of the most commonly used of these are briefly described below.

Prediction

In this method, data values are predicted from the values of previously gained ones. Only the difference between the predicted and the actual value is transmitted to the receiver, which can construct the true value from the previous values it has already received. This method is commonly used for motion compensation in video compression; *e.g.*, MPEG2.

Frequency based compression

These methods exploit the spatial and temporal attributes of images with regard to the sensitivities of the HVS. In sub-band video coding the different spatial-temporal frequency combinations are separated and coded according to their HVS sensitivities. Transform coding, which usually works in the spatial frequency domain employs energy compaction as described earlier. The most common transformation used is that of the Discrete Cosine Transform (DCT) (Ahmed et al., 1974).

[1]See section 2.2.5 in Chapter 2

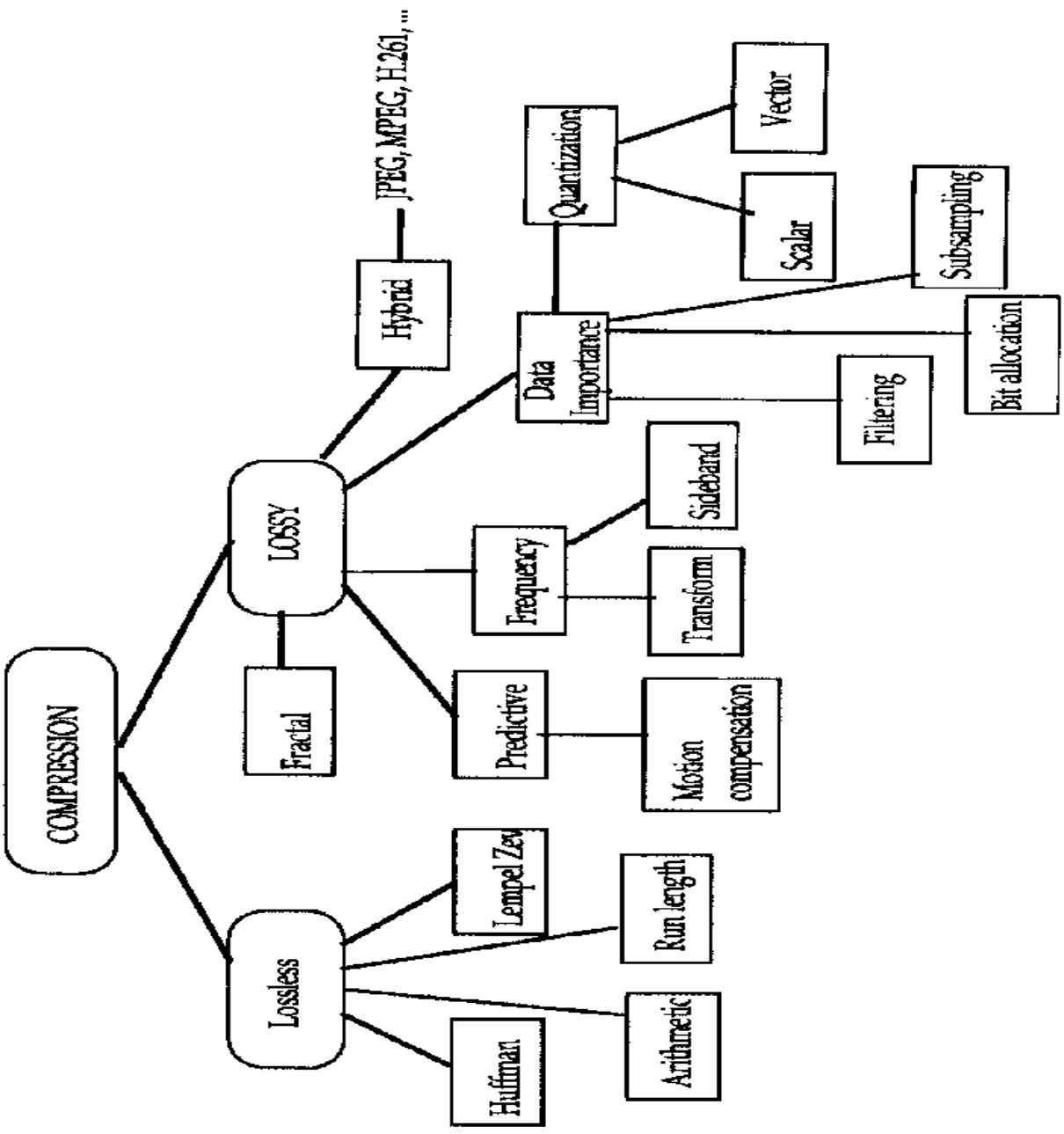


Figure 5.1: Topological diagram of common compression methods.

Feature based compression

Feature based compression centres on image features that are important for human image interpretation, such as edges, and considers the HVS response. This approach may be functionally similar to some frequency based methods, but performs them in the spatial domain. In filtering for example, details of an image beyond the discrimination of humans, may be removed. Bit allocation may be employed to allow more information to be encoded for important features, such as edges, and less to areas carrying less visual information, thus reducing the overall amount of data. An obvious form of compression can be achieved by removing every N^{th} data point from the image. This is called sub-sampling. Another form of compression in this category is that of quantisation, which is differentiated into scalar quantisation and vector quantisation. The former has already been briefly discussed simply as quantisation.

Compression is achieved using vector quantisation by breaking the original image into blocks, say 4x4, and mapping them to a set of code symbols (code book). A small number of bits representing the code book entry is transmitted. For a tutorial review see (Nasrabadi and King, 1988).

Fractal coding

In simple terms, a fractal is a geometric form which has self-similar, irregular details. Fractal image compression is closely related to vector quantisation coding, except that the image itself provides the codebook. Figure 5.2 shows a simplified diagram of the process used in the fractal compression algorithm used in this study. The image to be compressed, is first partitioned into non-overlapping sub-images (8×8 pixels in our case) called *range blocks* (R_i). The aim is then to find another sub-image which matches (by application of a similarity metric) this range block. This is achieved by dividing the image into sub-images called *domain blocks* (D_i), which are searched to find the best match, after a transformation, for each range block. This transformation (W_i) is defined by

$$W_i \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \end{bmatrix}, \tag{5.1}$$

where x and y are points in an Euclidean plane, which are linearly transformed by the matrix $\begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix}$, and are translated by the vector $\begin{bmatrix} e_i \\ f_i \end{bmatrix}$. This is called an (2-D) affine transformation; *i.e.* an object in the image plane undergoes any or all of rotation, scaling or translation[2].

Unlike the range blocks, the domain blocks can overlap. Even though the D_i can be of fixed size, the algorithm used here uses a quad-tree[3] (Brown and Ballard, 1982) method to partition the image, with a maximum domain block size of 16×16. This allows more flexibility in obtaining a good match between the R_i and D_i, and thus a better quality compressed image.

The compression comes about as only the parameters a_i, b_i, c_i, d_i, e_i and f_i in (5.1) (plus 2 others) have to be saved rather than every pixel value, for each range block. This process is very

[2] In the case of fractal image compression, W_i, must also be *contractive* (see ref (Fisher, 1993))

[3] A square in the image is broken up into 4 sub-squares, when it is not matched well enough by a larger domain.

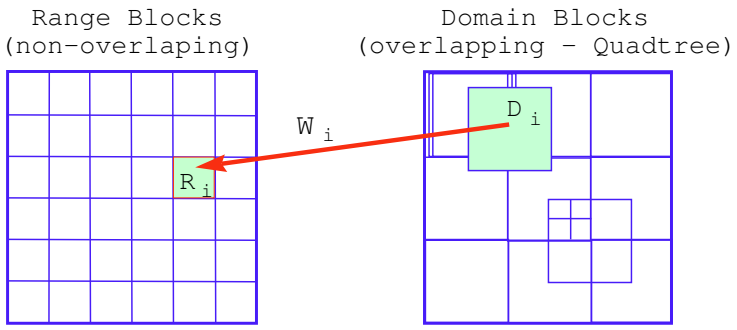


Figure 5.2: The encoding, with affine transformation, of the image blocks to form auto-codebook.

asymmetrical in that the time to compress an image is much longer than the time to de-compress an image. This is because the compression stage requires large searches to be carried out in order to match range and domain blocks. For further details see (Jaquin, 1993; Fisher, 1993).

Compression standards

With the greater integration of the world's telecommunications and computing resources it is necessary to lay down compression standards. As has just been discussed, there are numerous compression techniques in existence, but for the sake of standardisation, common usage of schemes must be decided upon. Standards need to be produced, but to be successful they also need to be generally adopted. The main contenders for grey scale or colour images and video compression standards are the ISO[4]/IEC[5] JPEG[6] (for still image) and MPEG[7] (for video). These have been briefly mentioned earlier, and MPEG is discussed in Chapter 6, while an overview of JPEG is given here.

JPEG is a hybrid method, which combines DCT coding with scalar quantisation and entropy coding, and can work in several modes. These include both lossy and lossless compression schemes. In this study, JPEG was used in the DCT based (lossy) mode. Shown in figure 5.3 is a basic block diagram of the JPEG coding method. The input image is partitioned into non-overlapping contiguous 8×8 pixel blocks. These image blocks are then processed as follows. The 64 image samples in each block are transformed, using the DCT, which is a near optimum[8] de-correlation and compaction step. As stated earlier, this transformation does not provide any compression; this is achieved in the next step - quantisation. Quantisation is achieved by dividing the DCT coefficients by the elements of the quantisation tables. Default tables are defined, but these can be specified by the user. The default tables take into account the HVS spatial frequency response in discarding information. Next the quantised data is further compressed by entropy encoding, and, unlike quantisation, this is achieved losslessly. As stated

[4]International Organisation for Standardisation
[5]International Electro-technical Commission
[6]Joint Photographic Experts Group
[7]Motion Picture Experts Group
[8]The Hotelling transform is the optimum method, but is much more expensive in computing time.

earlier, this method achieves compression by removing statistical redundancies in the data. This uses either of two schemes, Hoffman or Arithmetic encoding, with the former method requiring user-supplied tables.

Now the image has been compressed. To achieve de-compression, the reverse procedure is employed, as shown in the lower arm of figure 5.3. Unlike fractal compression, this process is symmetric, with the compression and de-compression times being equal. For a good description of the JPEG standard and its underlying principles see, Pennebaker and Mitchell (1993).

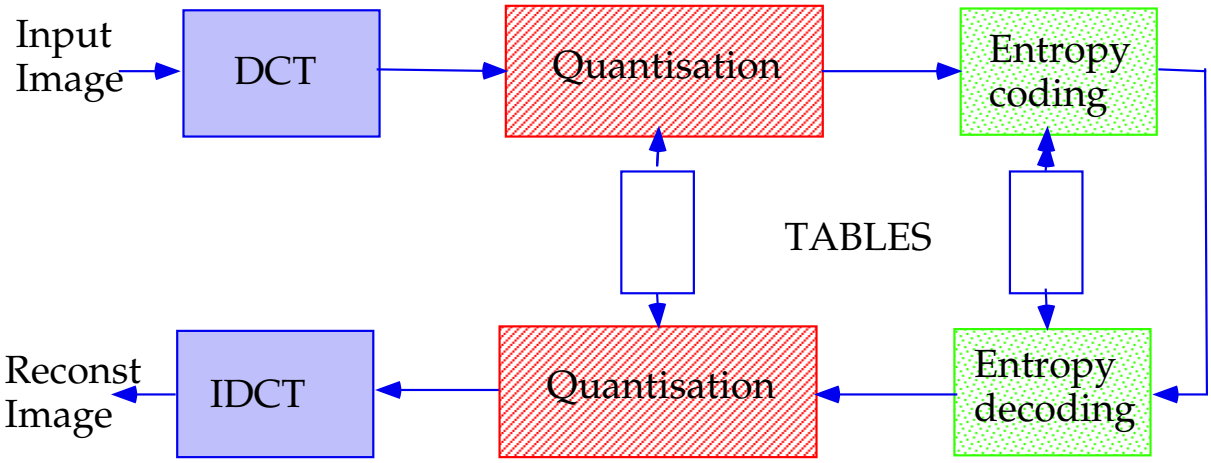


Figure 5.3: Block diagram of the JPEG compression & decompression scheme.

5.2 Methodology

The experiment required subjects to locate single targets quasi-randomly located in a number of different images, with variation in the target size and in the contrast between target and background. The images used had been compressed and then decompressed at a variety of compression ratios using two different compression methods. The compression methods used were implementations of the JPEG algorithm and the fractal transform. Response times, proportion of targets acquired, and observer confidence levels were collected for each trial. These were analysed against target size, contrast, compression ratio and compression scheme using an analysis-of-variance (ANOVA) (Hines and Montgomery, 1980a) approach.

5.2.1 Estimation of Sample Size

It is necessary to determine the number of observations required to guarantee a specified level of significance and the power of a statistical test used in an experimental design. In these experiments we are comparing the means for various treatments (combinations of parameter values). The null hypothesis is

$$H_0 : \mu_1 = \mu_0;$$

while the alternative hypothesis is

$$H_a : \mu_1 \neq \mu_0.$$

The methodology used for estimating statistical power is outlined in Appendix C. Using an estimate of the variance from a related experiment, carried out on the effects of zooming (Woodruff and Newsam, 1994), the estimate for sample size for each compression method was calculated as 353.

5.2.2 Pre-experiment

In order for an experiment of this type to be run efficiently, the stimulus characteristics have to lie within the perceptual range of the HVS. They must also lie within a range such that a change in stimulus properties invokes a change in response; *i.e.* the stimuli should be on the approximately linear portion of the psychometric function[9]. As an extreme example, if the contrast of the targets were all too high, then little discrimination on the effects of target detection would be achieved, as the hit-rate would remain constant and high. Conversely, if the contrast was too low (below threshold) the hit-rate would be zero.

In order to estimate these limits for the target stimuli, I carried out a pilot experiment. Since the luminance threshold is a function of target area (Blackwell, 1946; Lukis and Budrikis, 1982), when viewing small or low contrast targets, the size-contrast product (ca_t) was used as an independent variable in this pilot study.

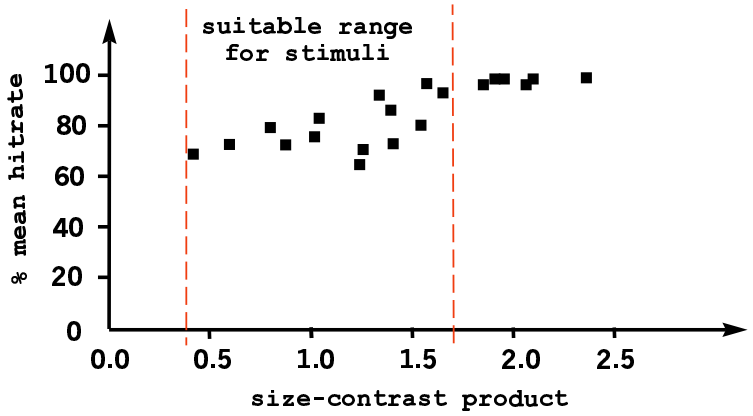


Figure 5.4: Finding the range for the size-contrast product of the targets. The lower red line indicates the threshold, below which, no detections were obtained. The upper red line indicates the threshold, above which, 100 % detection was obtained.

The software that I wrote to perform the main experiment had several test modes. One of these restricted the compression ratios to maximum and zero (no compression) and allowed the operator to choose the compression scheme. Under this mode, I ran the pilot experiment and viewed 600 stimuli over a few sessions. The maximum difficulty trials (with high compression)

[9]See Chapter 3 section 3.1.1

determined the lower bound for the stimuli, while the lower difficulty trials (with no compression) determined the upper bound for the stimuli. I chose fractal compression as my experience suggested it interfered with target detection at maximum compression more than JPEG. (This was borne out by the experimental results.) The results are summarised in figure 5.4, where the appropriate range for ca_{t} is indicated. The actual values determined for the stimuli parameters in the main experiments are given in Appendix D.

5.2.3 Stimuli

Images from a rural Australian dry temperate area were obtained with a thermal imager operating in the 8-12 micron band. Noise reduction by frame integration over 256 frames was used to improve the grey level resolution of the imagery. These images were 512×512 pixels in size, each pixel being represented by a single byte. Four different scenes, as shown in figure 5.5, were used as backgrounds. Each scene was a ground-to-ground view from a moderately elevated position, resulting in frames consisting of little or no sky.

The pixel values of the background images were linearly transformed so as to have the same mean and standard deviation. The mean grey-level was set to 100. The basis for the selection of the mean grey-level was my perception of the optimal luminance dynamic range for the displayed scenes, given the monitor calibration status as described in section 5.2.4 on the next page. Test images were characterised by vectors of the form [compression method (cm) compression ratio (cr), target size (s_t), target contrast (c), background, region, location], which were generated as follows:

(1) For each background, a rectangular region for target insertion was specified;

(2) Next a target centre was randomly selected from within the chosen region;

(3) A random pair (s_t, c) from within the allowed domain of target size and target contrast was chosen and a target with these characteristics inserted at the chosen centre;

(4) A compression ratio $cr \in 0, 5, 10, 15, 20, 25, 30, 35$ was then selected randomly, subject to the constraint that each background be used equally often at each compression ratio. The test image was then compressed to this degree by both the fractal and JPEG algorithms. Finally, the resulting two files were decompressed back to the original 512×512 size and saved as test images. Example test images are shown in figure 5.6 for the same image for both JPEG and fractally compressed images.

The regions described in item (1) were of size 20×5 pixels, and some of these were located in improbable areas (*e.g.*, sky) so as to encourage a full search at each trial. In order to reduce any learning affects, 15 such regions were manually pre-selected for each background image. This was done so that the spatial distribution within each background image, and the distribution of background types within each region were similar for all the background images. Twelve of the 15 regions were then randomly selected by the software, during processing of the test images. Within the subset of test images with a given background, targets were systematically

and uniformly allocated across regions. Once the region had been selected, a centre for the target was randomly chosen within the region, using continuous coordinates, so that the target could arbitrarily overlap pixel boundaries. This is shown graphically in figure 5.7.

Targets of selected size and contrast were embedded in these scenes, using an insertion procedure described in Woodruff and Newsam (1994), which is detailed in Appendix E. Details of the choice of size and contrast are given in Appendix D. Briefly, a base target size of 4 pixels wide and 9 pixels high was specified, where 1 pixel subtended a visual angle of 7.22×10^{-2} degrees. Based on a pilot study, upper and lower bounds were placed on target area and contrast, while the product of contrast and area was constrained, since the simultaneous combination of minimal contrast and minimal area gave too low a detection rate, with converse effects for maximal values (see section 5.2.2). Within the allowed range of contrast and size values, these variables were distributed as uniform random variables.

To achieve ideal target insertion, the background images should be deconvolved to remove the effects of the point spread function[10] (PSF) of the imaging sensor system. Then the target could be inserted and the whole image re-convolved with the original psf. However, since deconvolution is notoriously fraught with problems, this approach was avoided. Instead, the following approximate approach was adopted. The method detailed in Appendix E was used with the following smooth zooming modification. A 32×32 pixel region co-centred with the target location was "cut out" from the background image and smooth-zoomed (Newsam, 1993) by a factor of 4. The target was then inserted in the centre of the region and convolved with a PSF which was consistent with the sensor that produced the background image. A sub-region of 64×64 pixels co-centred with the existing region, was then removed and pixel averaged down by a factor of 4 (dezoomed) and re-inserted in the background image. This smooth zooming procedure is equivalent to performing a bi-cubic spline interpolation between pixels in the excised image region. This was done to obtain PSF blurring at the sub-pixel level.

5.2.4 Apparatus

The experiment was controlled by a 486 33 Mhz DX personal computer. Images were presented on an Electrohome 1719X high quality monochrome television monitor from a Matrox PIP-1024 image digitising and display card. The relation between screen luminance, L, in candelas per square metre ($\mathrm{cd/m^2}$), and pixel grey-level value, g, was closely approximated by the quadratic function $L = 0.00243g2 + 0.10745g + 0.59044$, with $g \in \{0, \cdots, 255\}, (R2 = 0.999)$. A hood was placed over the screen to constrain the viewing distance to 500 mm and block out ambient light.

5.2.5 Subjects

Subjects were all volunteers from within the age group 23 - 38 years, having at least 6/6 vision. One subject was female. Ten subjects were used in Experiment 1 and six in Experiment 2. Three of the subjects from Experiment 1 also served as subjects for Experiment 2. None of the

[10] See section 2.2.4 in Chapter 2 for a definition of PSF.

subjects had previously served as subjects for similar psychophysical experiments. However, all subjects were professional engineers or scientists with familiarity with imaging systems.

5.2.6 Procedure

Experimentation consisted of three phases: first a pilot study to determine appropriate ranges for size and contrast values (which has already been discussed); secondly, experiment 1, in which the test-retest reliability of the procedure was determined; and, finally, experiment 2, which was the actual compression study. In Experiment 1, half of the stimuli - equally distributed between JPEG compressed and fractal compressed - were presented twice to each subject in random order. In Experiment 2 each stimulus was presented exactly once to each observer in random order. In all cases the orders of presentations for subjects were uncorrelated.

The presentation procedure - which was the same in both experiments - was as follows:-

(i) Both experiments consisted of 768 presentations, referred to here as trials. The entire set of trials was completed over several sessions, which were of a duration of about 30 minutes. The order of presentation of stimuli to each observer was pre-calculated at the commencement of that observer's first session, and was random. To minimise learning affects, the initial session for each subject was treated as a training period. That is, the results of this session were deleted and the experiment was re-initialised in the following session. This then became the real start of the experiment. All stimuli were presented exactly twice in Experiment 1 (*i.e.* half the set was used), and exactly once in Experiment 2. Data for each experiment were collected over a period of approximately 5 - 7 days for each observer, with subjects typically being able to complete the full set of trials for an experiment in 3 - 4 sessions.

(ii) Each trial consisted of the following sequence:-

  (a) stimulus presentation,

  (b) observer search for a maximum of 9 seconds (with search normally being terminated by the observer depressing the spacebar of a standard computer keyboard),

  (c) computer presentation of a centrally-located cursor which the observer then moved via a mouse to the putative target position (or to a random location if no target was found) and clicked the mouse button,

  (d) entry of a confidence level in one of three categories (certain, not sure but about 50% confident, pure guess/no idea).

The screen was then blanked with a uniform grey at the mean grey level of all images, and a new search scene loaded into the framestore. The subject could pause the experimental session or terminate the session on completion of any trial in which they were sure of the target location.

Experiment 1

Each subject was presented with two identical sets of 384 images, but the order of presentation of each image was random and different for each subject and each run. Each image in a set represented a unique stimulus treatment; *i.e.* each image represented a unique combination of: the 4 background images, the 12 target size-contrast levels, and the 8 compression ratios. For each subject, a target hit (detection) or miss was recorded, along with the search time in case of a hit. After each image presentation, the subject entered a confidence rating. This was used to verify that a hit had occurred (discussed in section 5.3 on page 105).

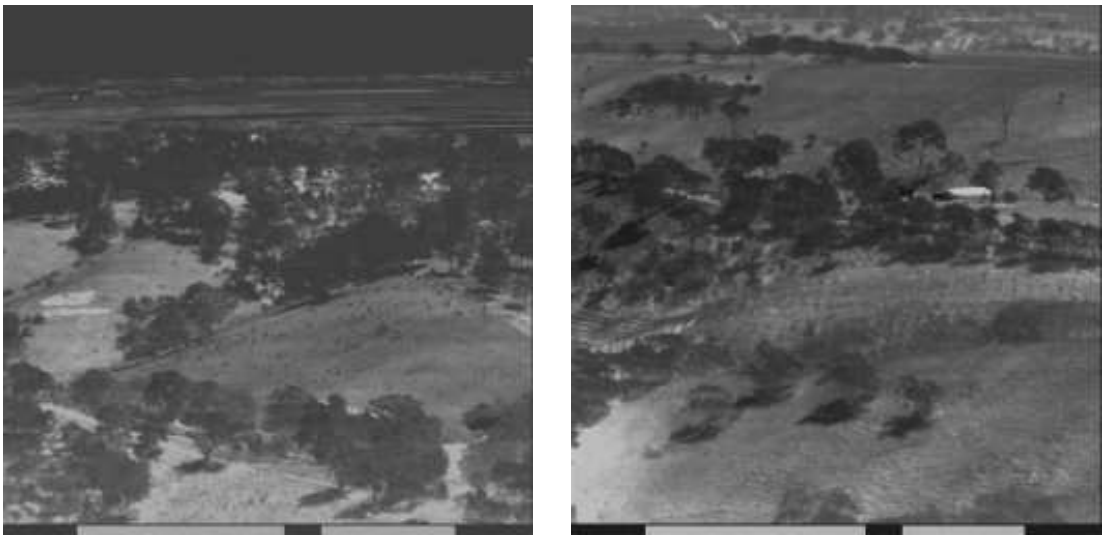
The subject's responses were averaged over subjects to produce a mean search time (MST) and a mean hit-rate (MHR) for each test image displayed. In order to determine if any significant learning effect had occurred, a paired t-test was applied to each pair of response data (1st and 2nd attempts), for both mean hit-rate and mean search time. A pair-wise correlation analysis was applied to the 1st and 2nd run data for both mean search time and mean hit-rate results.

Experiment 2

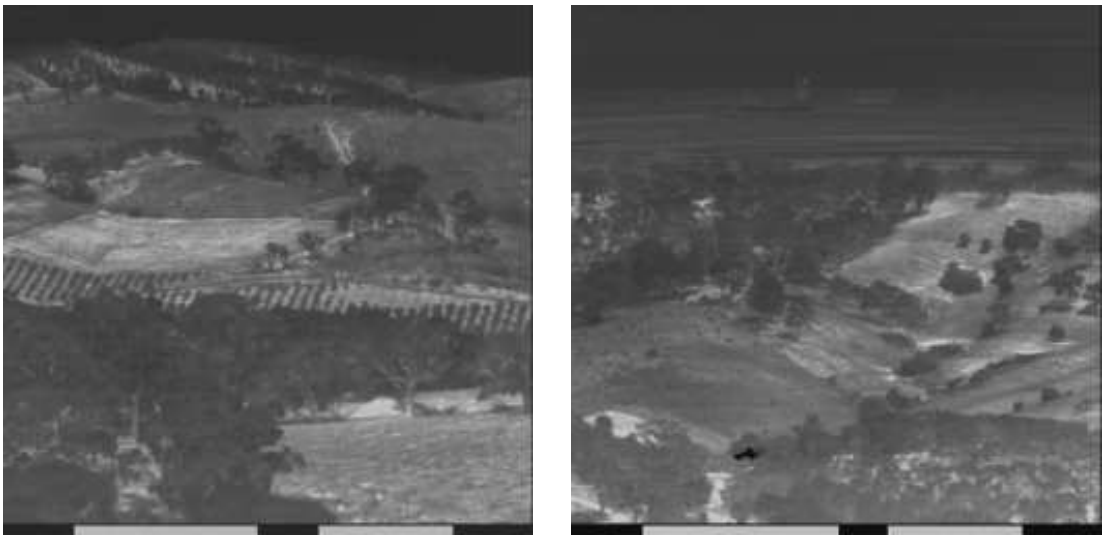
As mentioned previously, this experiment was performed to assess the effects of both compression method and ratio on target detection. Experiment 2 conformed to a mixed random and fixed effects design with blocking (Hines and Montgomery, 1980b). The blocking variables were Subject, background and within-image location. Compression method (cm) and compression ratio (cr) were fixed effects - *i.e.* they had selected discrete values, while Subjects were treated as a block variable, with each subject seeing all of the test images.

In this second experiment, six subjects participated, of whom three had participated in experiment 1. Again, 768 images was presented to each subject in a different random order. However, this time, each image represented a unique treatment, with the extra factor of compression method being included in the factor combinations; *i.e.* each image represented a unique combination of: the 4 background images, the 12 target size-contrast levels, the 8 compression ratios, and the 2 compression methods. As in Experiment 1, a hit or miss was recorded for each subject at each image presentation, with the search time also being recorded for a hit. A confidence level was also recorded by the software.

A statistical analysis of the experiment was obtained by performing a two-way classification (cm, cr) ANOVA.

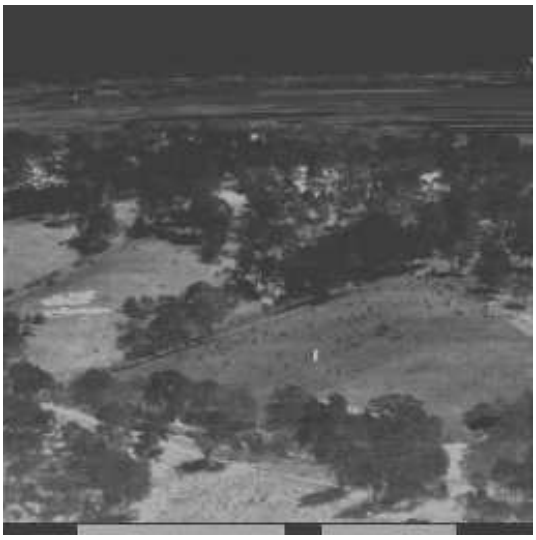


(a) (b)



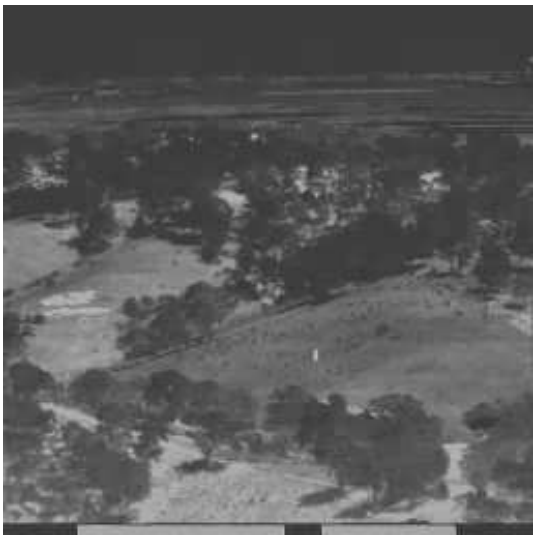
(c) (d)

Figure 5.5: The four original Infrared background images.

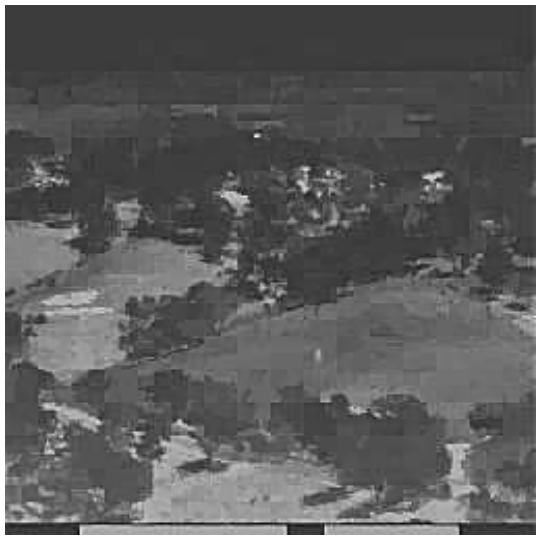


(a) JPEG at 10:1 compression ratio.

(b) JPEG at 35:1 compression ratio.



(c) Fractal at 10:1 compression ratio.



(d) Fractal at 35:1 compression ratio.

Figure 5.6: Examples of Compressed and then decompressed images.

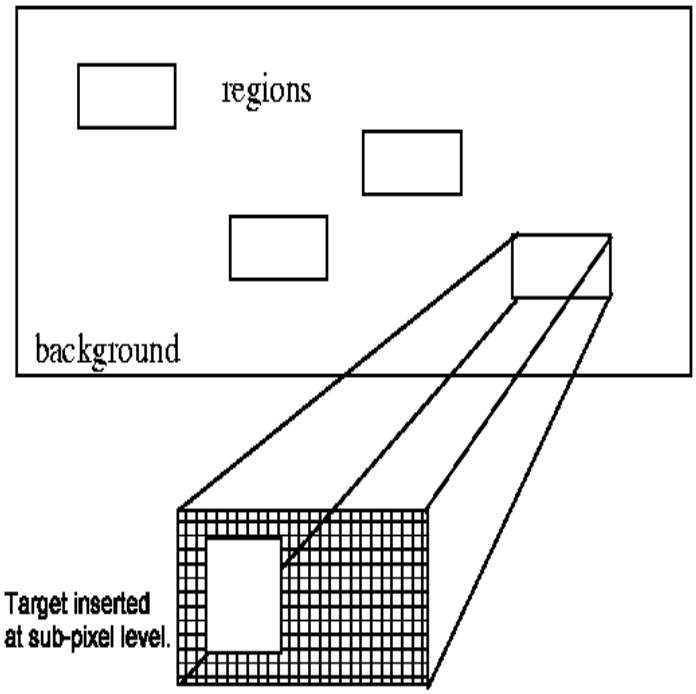


Figure 5.7: Procedure for sub-pixel insertion of targets.

5.3 Results & Discussion

The performance measures obtained on each trial were search time and acquisition (hit or miss). Since the mean acquisition level at different compression ratios varied from near 100%, down to approximately 50%, the use of search time as a measure of target detectability is unsound, for reasons which are discussed by Woodruff & Newsam (1994) and by Ewing & Woodruff (1996). However, a brief explanation is now given. The hit-rate indicates the degree of difficulty involved in target detection for the experimental subjects. At different hit-rates, say 60% compared to 50%, different populations of targets are involved, and the target population found with the lower hit-rate is harder to detect than the target population found with the higher hit-rate. Therefore, at the different hit-rates, different biases are introduced into the search time data. Another way of stating the above is to say that the subjects are operating at different points on their receiver operating characteristic curves[11] for different hit-rates.

The confidence level was used to verify whether a target had been truly acquired. For example, if a hypothetical target had been correctly located, but the confidence level indicated the subject had no confidence in this detection, then a "miss" would have been recorded. The confidence level was not explicitly used in the analysis.

The following probability of detection measure (mean hit-rate) was used:

$$p_{d_i} = \frac{1}{N}\sum_j W_{ij}, \text{ where } W_{ij} = \begin{cases} 0 & \text{miss} \\ 1 & \text{hit} \end{cases}, \tag{5.2}$$

for the i^{th} treatment and j^{th} subject; where N = number of subjects.

5.3.1 Experiment 1 - Reliability Analysis

From Experiment 1 only reliability data were sought. For each subject, an ordered pair is available for each of the 384 stimuli - not found (0) or found (1) on the first and second presentations of that stimulus. The pairwise correlation between first and second presentations over the complete set of response data was calculated and is presented in table 5.1. Collapsing data over the set of subjects to give the number of targets correctly located (hits), and then computing the correlation, gave a group test-retest reliability, which is also given in table 5.1. The high correlations obtained establish that the procedure is highly reliable - particularly when group means are used. In addition, using the number of hits for each stimulus on first and second presentations, a t-test of the difference in the mean hit-rate was computed. However, no significant difference was obtained between first and second presentations (see table 5.2). This suggests that there was no significant learning effect operating.

Nevertheless, since the hit-rate from the first trial was very highly correlated with that from the second trial, it was valid to use search time as a performance measure for test-retest reliability assessment. We have therefore done this and obtained the following results:-

[11] Refer to section 3.2.2 in Chapter 3.

Measure	Corr	95% Lower	95% Upper	P-Value
MHR	0.971	0.965	0.977	< 0.0001
MST	0.968	0.961	0.974	< 0.0001

Table 5.1: Correlation analysis between 1st and 2nd attempts in experiment 1, for both the mean hit-rate (MHR) and the mean search time (MST).

Measure	Mean Diff	DF	t-Value	P-Value
MHR	0.003	383	0.882	0.3782
MST	0.154	383	5.181	< 0.0001

Table 5.2: Paired t-test between 1st and 2nd attempts in experiment 1. The Mean Difference for MHR is in probability units and for MST it is in seconds. The mean hit-rate was a stable measure of detection performance for data obtained for 1st and 2nd sets of trials of the same stimuli, while mean search time was not.

(i) Test-retest reliability is 96.8%, $p < 0.0001$;

(ii) Difference in mean search rate for items acquired in both attempts was 0.154 seconds, $p < 0.0001$.

This again indicates that the procedure is highly reliable. A small, but statistically significant, decrease in mean search time was obtained for the second presentation of the stimulus set when compared to that of the first presentation. Coupled with the fact that hit-rate (or task accuracy) did not improve, this suggests that the small decrease in search time was due to the subject's familiarisation with the experimental procedure, rather than learning in the actual visual task.

5.3.2 Experiment 2 - Effects of Compression

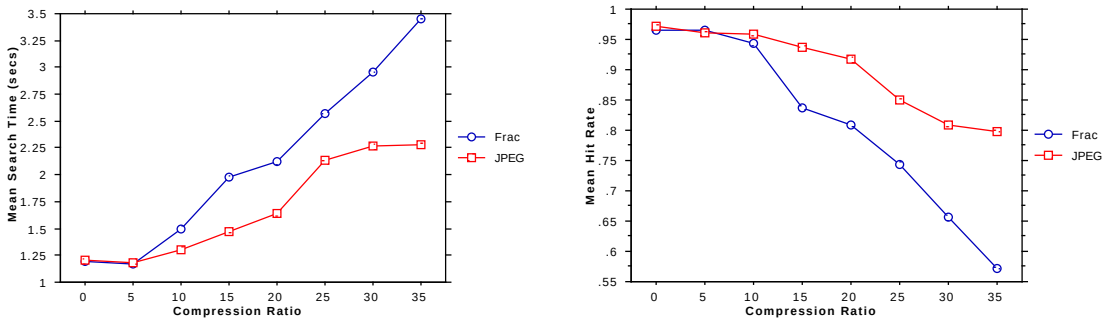
The results of experiment 2 are graphically presented in figure 5.8 for both dependent variables;*i.e.* MST (figure 5.8(a)) and MHR (figure 5.8(b)). Figure 5.8 indicates that better overall observer performance was obtained when viewing JPEG compressed images rather than fractally compressed images, and that there is little effect on performance for compression ratios less than 10:1. This is confirmed by the ANOVA, which is discussed in the remainder of this section.

The results of the ANOVA using MHR as the dependent variable are shown in table 5.3.

Source	DF	Sum of Squares	Mean Square	F-Ratio	P-Value
CM	1	1.524	1.524	18.233	< 0.0001
CR	7	7.894	1.128	13.493	< 0.0001
CM*CR	7	1.068	0.153	1.826	0.0794
Error	752	62.851	0.084	-	-

Table 5.3: ANOVA table for mean hit-rate (MHR) as the dependent variable. The p-values calculated for an α level of 0.05.

Perusal of table 5.3 with MHR as the dependent variable shows that the main factors



(a) Search time vs compression ratio.

(b) Hit-rate vs compression ratio.

Figure 5.8: Interaction of compression type with compression ratio.

gave highly significant effects (cm: $p < 0.0001$, cr: $p < 0.0001$), but the interaction between compression method and compression ratio failed to reach significance ($p = 0.0794$).

In order to investigate this last point further, a post-hoc analysis of the cm*cr interaction was done by applying the F-test to the variance ratios $\frac{\text{varJPEG}}{\text{varFRAC}}$ at each compression ratio with the results shown in table 5.4.

Comp Ratio	Num DF	Den DF	F-Ratio	P-Value
Total	383	383	0.515	0.0001
0	47	47	0.920	0.7774
5	47	47	1.015	0.9599
10	47	47	0.934	0.8147
15	47	47	0.390	0.0016*
20	47	47	0.442	0.0060*
25	47	47	0.580	0.0652
30	47	47	0.622	0.1068
35	47	47	0.519	0.0265*

Table 5.4: Post hoc analysis of interaction cm*cr for mean hit-rate.

From table 5.4, it can be seen that the p-values for all the data (total) agrees with the p-values obtained for the effect of compression method in the ANOVA tables. Table 5.4 indicates that significant cm*cr interactions exist for $cr \in \{15, 20, 35\}$.

Effect means for cm are shown in table 5.5 for MHR.

CM	Count	Mean	Std. Dev.	Std. Err.
Frac	384	0.811	0.352	0.018
JPEG	384	0.900	0.252	0.013

Table 5.5: Mean effects due to compression method for MHR.

It was found that the mean hit-rate when viewing JPEG compressed images was an average of 11% (8.0% - 12.8%) higher (over the whole range of cr), than that obtained when displaying fractal compressed images.

To test for the significance of the difference in the cell means for JPEG versus fractal compression at the various compression ratios, paired t-tests were performed. These are depicted in table 5.6 for MHR as dependent variable. These data indicate that, for $cr > 10$, there were significant differences in the mean subjective responses associated with the compression method used. Using MST, a just significant difference was detected at $cr = 10$, which was not significant using MHR, but the latter has detected differences at greater significance levels for all higher compression ratios.

Comp Ratio	Mean diff	DF	t-Value	P-Value
Total	0.089	383	7.077	0.0001
0	0.007	47	0.984	0.3299
5	-0.004	47	-0.443	0.6595
10	0.014	47	1.671	0.1013
15	0.101	47	2.780	0.0078*
20	0.108	47	2.969	0.0047*
25	0.108	47	3.017	0.0041*
30	0.153	47	3.165	0.0027*
35	0.225	47	4.156	0.0001*

Table 5.6: Paired t-test of mean hit-rate scores for JPEG - fractal.

5.4 Conclusions

The results from the reliability analysis show that the technique used is highly reliable, particularly when group rather than individual data is used. A learning effect was noted, but of such a size that it did not show up in the less sensitive (but, more robust) "hit-rate" performance measure used in the analysis of the main experimental data.

The targets used in this study occupied from 0.01 to 0.04% of the total image, and were of moderate contrast (0.2 - 0.6). They therefore constitute only a very small component of the total energy of any image. The results of experiment 2 clearly show that acquisition of such targets, following compression by compression ratios of 15 or greater and subsequent decompression, is significantly better if JPEG compression is used rather than fractal compression. At low compression ratios ($cr < 10$) there is no significant difference in performance.

There is not any other comparable work in the literature with which to compare the results of this chapter. There is a gap in the literature on the effects of image compression on target acquisition in humans. There has been reported some work on the effects of image compression on subjective image quality, but of the two compression methods used here, there appears to be only limited work on JPEG compression and image quality (Malo et al., 1997; Fuhrmann et al., 1995; Kostas et al., 1993; Watson, 1993). However, this work on the effects of JPEG compression only considers subjective ratings of aesthetic image quality. An exception is the work done by Kostas, Sullivan & Ansari (1993), which considers the effects of (JPEG) compression on (medical) visual task performance, but even here only the subjective ratings of expert users (radiologists) are used.

Chapter 6

Studies on the Effects of Video Compression on Target Recognition

Summary: *Earlier work studied the effect of still image compression on target detection. This chapter continues in a similar vein, but examines the effects of video compression on target recognition. A set of video compression experiments were performed, which required each of 10 observers to recognise ships in 512 randomly presented video sequences. The sequences, which lasted for five seconds, had controlled levels of contrast and multiplicative noise, and were compressed and de-compressed at a variety of compression levels using MPEG-2 encoding under standard settings. Response timesand recognition accuracy were collected for each trial. These were analysed using analysis-of-variance (ANOVA) techniques.*

6.1 Introduction

The expected proliferation of digital imaging systems in the near future will be accompanied by the widespread use of image compression to reduce transmission and storage requirements. Lossless[1] image compression can be attained, but this achieves only the relatively low compression ratios of about 5:1 or less. If higher compression ratios are needed, then lossy compression is required. Much military and protective surveillance is concerned with the detection and recognition of relatively small, low contrast targets. If surveillance imagery is to be transmitted from point to point, or if it is to be stored for later analysis, then lossy compression schemes may be required. Therefore, when considering an imaging system for a well defined visual task, such as surveillance, it is important to first assess to what extent image compression is likely to affect user performance on the given tasks. If possible, compression schemes ought to be tuned to the specific task(s) which will be carried out with these images.

In evaluating image compression techniques, many factors come into play, including implementation complexity, real-time processing considerations, compression ratio, *etc.* Nevertheless the over-riding concern is the quality of the reconstructed image. Hence there is a great need to

[1] See section 5.1.1 in Chapter 5

systematically evaluate image compression algorithms with respect to their degradation of image quality. However, the literature remains silent on the issue of evaluating human visual task performance using MPEG compressed video. Of the work evaluating MPEG (including MPEG-2) video compression, only a subjective rating method has been used (Hidaka and Ozawa, 1993). In fact only the sunbjective rating of image quality is specified in the MPEG standards, which are discussed in section 6.2.

This work assessed the affects of video compression on target recognition and identification by observers. In particular, the main aim was to quantify the degradation in target acquisition performance when viewing short sequences of video imagery which had been compressed under the new, MPEG-2 standard, which is discussed in the next section.

6.2 MPEG Video Coding

As this chapter is concerned with the effects of MPEG-2 compression on human visual performance, an overview, with some background information on MPEG compression, is now given. MPEG compression was briefly discussed in Chapter 5 in the context of compression methods. For more details, see the actual standard specification (ISO/IEC 13818-2), though this is hard going, or see a text, such as that by Tekalp (1995), which covers the theory and various video coding standards, or the paper by Chen (1995) for a review of digital video compression standards.

Commercial television is undergoing a transition to digital processing and transmission, but currently is analogue. Analogue video standards define the number of frames per second (29.97 for NTSC[2] and 25 for PAL[3]) and the number of lines per frame (525 for NTSC and 625 for PAL). Video signals also contain a blanked portion that is used for synchronisation but not displayed, so not all lines nor all parts of each line contain active video. To convert to digital, the analogue video signal is sampled along each of the active video lines (486 for NTSC and 576 for PAL). A common rate is 13.5 MHz, defined in CCIR- 601[4] and used for D-1[5] and the new Video-CD format for compressed video. The MPEG standards, which are defined for digital video, are now discussed.

MPEG, which stands for Moving Pictures Experts Group, is a joint committee of the International Standardisation Organisation (ISO) and International Electro-technical Commission (IEC). It has been responsible for the MPEG-1 and MPEG-2 standards in the past and has just released the final draft on the MPEG-4 (ISO/IEC, 1998) standard, which is due to be formally ratified in January 1999 as ISO/IEC 14496. MPEG standards are generic and universal in the sense that they merely specify a compressed bitstream syntax. This, in effect, unambiguously defines the de-compression process. However, the standard does leave room for smart implementations of the encoder, compression algorithm, and the decoder. There are three main parts of

[2]The North American television standard.

[3]The European and Australian (in slightly modified form) standard.

[4]Digital television standard promulgated previously by the CCITT and now the International Telecommunications Union (ITU).

[5]A digital video format standard for very high quality (studio) recording and play-back.

the MPEG-1 and MPEG-2 specifications, namely, Systems, Video and Audio. The Video part defines the syntax and semantics of the compressed video bitstream. The Audio part defines the same for the audio bitstream, while the Systems part addresses the problem of multiplexing the audio and video streams into a single system stream with all the necessary timing information. The timing information is necessary to synchronise the playback of the stream by the decoder, without any overflow and underflow of the decoder buffers. Additionally, the MPEG-2 specification consists of a fourth part, called DSMCC, which defines a set of protocols for the retrieval and storage of MPEG data from and to a digital storage medium.

MPEG-1 and MPEG-2 have different feature sets, targeting different applications. The MPEG-1 standard addresses desktop multimedia applications such as storage and retrieval of data from CD-ROMS at bit-rates close to 1.5 megabits-per-second (Mbps). The quality of MPEG-1 video is usually better than VHS quality video (see figure 6.1). Other applications include video-conferencing, electronic publishing, games, video-mail and video-phone. At low bit-rates, before compression, video is usually decimated[6] to MPEG "standard input format" (SIF) resolution, 360×243. The reason for this is shown in figure 6.1, which depicts two curves: the CClR-601 curve corresponds to compressing video at full input resolution (720×486); the other corresponds to compressing video at SIF input resolution. Figure 6.1, indicates that better quality, relative to standard play-back technologies, is achieved using SIF input for lower bit-rates, but this advantage is soon lost for higher bit-rates ($>\approx$ 3Mbps).

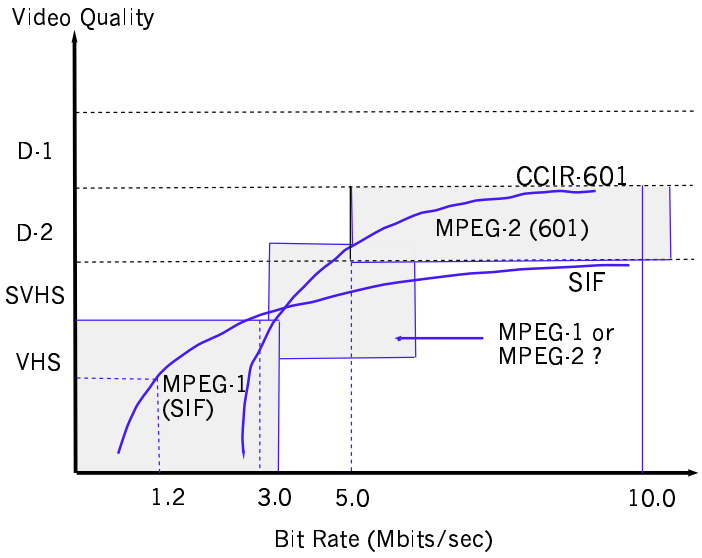


Figure 6.1: Video quality for MPEG

MPEG-2 compresses full resolution CCIR 601 video at bit-rates between 4-15 Mbps for a whole range of telecommunication applications that need broadcast quality video. Its applications include among others cable television, and high definition television (HDTV). MPEG-2 was designed for the higher quality, high bit-rate applications and MPEG-1 for lower bit-rates.

[6] That is, every n^{th} pixel is dropped. In this case $n = 2$.

However, there are no firm constraints in either algorithm, and it is possible to run MPEG-1 video at a very high rate, or MPEG-2 video at a very low rate.

As the technology is improving rapidly, there is a continuing demand to achieve higher compression without seriously compromising the quality. That has been a major focus for the MPEG video experts as they currently develop the MPEG-4 standard for low bit-rate, content based coding solutions for interactive audio-video applications.

6.2.1 MPEG-2

MPEG-2 video is a generic method for compressed representation of video sequences, using a common coding syntax, defined in the document ISO/IEC 13818 Part 2 by the ISO and the IEC. This standard was developed in collaboration with the International Telecommunications Union (ITU) who promulgate it as Recommendation H.262.

The MPEG-2 concept is similar to MPEG-1, but includes extensions to cover a wider range of applications, with the essential difference between them being the incorporation in MPEG-2 of field[7], rather than frame-based processing, in a technique called "field-based motion prediction".

Several other more subtle enhancements (*e.g.*, 10-bit DCT DC precision, non-linear quantisation, improved mismatch control) are included, which have a noticeable improvement on coding efficiency, even for progressive video. Other key features of MPEG-2 are the scalable extensions. These permit the division of a continuous video signal into two or more coded bit streams representing the video at different resolutions, picture quality (*i.e.* signal-to-noise ratio), or picture rates.

6.2.2 MPEG Algorithms

The basic idea behind MPEG video compression is to remove spatial redundancy within a video frame and temporal redundancy between video frames. As in JPEG[8], the standard for still image compression, DCT-based (Discrete Cosine Transform) compression is used to reduce spatial redundancy. Motion-compensation is used to exploit temporal redundancy. The images in a video stream usually do not change much within small time intervals. The idea of motion-compensation is to encode a video frame based on other video frames temporally close to it.

At the highest level of the hierarchy, the video bitstream consists of video sequences;*i.e.* sequences of pictures. MPEG-1 allows for only progressive sequences, while MPEG-2 allows for both progressive and interlaced sequences. Each video sequence consists of a variable number of groups of pictures (GOP), where a GOP contains a variable number of pictures.

A picture can either be a frame picture or a field picture. In a frame picture, the two fields are coded together to form a frame, while field picture is a coded version of an individual field. Pictures can be either of frame type or field type in MPEG-2, while MPEG-1 allows only frame

[7]Each frame consists of two interlaced fields, so the field rate is twice the frame rate.
[8]See section 5.1.1 in Chapter 5

pictures. Pictures can be categorised into three main types, based on their compression schemes.

- I or Intra pictures;
- **P** or Predicted pictures;
- **B** or Bi-directional pictures.

I pictures are coded by themselves (hence the name, Intra). The coding technique for these pictures falls in the category of transform coding. Each picture is divided into 8×8 non-overlapping pixel blocks. Four of these blocks are additionally arranged into a bigger block of size 16×16, called a macroblock. The Discrete Cosine Transform is applied to each 8×8 block individually. The transform exploits the spatial correlation of the pixels by converting them to a set of independent coefficients. The low frequency coefficients contain more energy than the high frequency ones. These coefficients are quantised, employing a quantisation matrix, as discussed in Chapter 5 for the JPEG algorithm. Quantisation is the only lossy part of the whole compression algorithm other than sub-sampling. The quantisation process also helps in rate control, *i.e.* allowing the encoder to output bit-streams at a specified bit-rate.

The DCT coefficients are coded employing a combination of two special coding schemes: Run Length and Huffman (entropy) coding. Each block of coefficients is scanned in a zigzag pattern to create a 1-D sequence. MPEG-2 can additionally provide a different scan pattern as an alternative. The resulting 1-D sequence usually contains a large number of zeros, due to the lowpass nature of the DCT spectrum and the quantisation process. The non-zero DCT coefficients are allotted a variable length code from a lookup table. This is done in such a manner that a highly probable combination gets a code with fewer bits, while the unlikely ones get longer codes. Adopting this lossless coding technique, the total number of bits is kept down. However, since spatial redundancy is limited, the **I** pictures provide only moderate compression. These pictures provide important hooks for random access into the digital bitstream for editing purposes. The frequency of **I** pictures is normally once every 12 to 15 frames.

The **P** and **B** pictures are where MPEG derives its maximum compression efficiency. It does that by a technique called motion compensation (MC) based prediction, which exploits temporal redundancy. Since frames are closely related, it is assumed that a current picture can be modelled as a translation of the picture at a previous time. It is possible then to accurately represent or"predict" the data of one frame based on the data of a previous frame, provided the translation is estimated. The process of prediction helps in the reduction of bits by a huge amount. In **P** pictures, each 16×16 pixel sized macroblock is predicted from a macroblock of a previously encoded **I** picture. Since, frames are snapshots in time of a moving object, the macroblocks in the two frames may not be co-sited,*i.e.* correspond to the same spatial location. Hence, a search is conducted in the **I** frame to find the macroblock which closely matches the macroblock under consideration in the **P** frame. The difference between the two macroblocks is the prediction error. This error can be coded before, or after DCT transformation. The DCT of the error results in few high frequency coefficients, which after the quantisation process require a small number of bits for representation.

The quantisation matrices for the prediction error blocks are different from those used in intra blocks, due to the distinct nature of their frequency spectra. The displacements in the horizontal and vertical directions of the best match macroblock from the co-sited macroblock are called motion vectors. The motion vectors represent the translation of the picture blocks between frames. These vectors are obviously needed for reconstruction and are differentially coded in the bitstream. Differential coding is used because it reduces the total bit requirement by transmitting the difference between the motion vectors of consecutive frames. The compression efficiency and the quality of the reconstructed video depends on the accuracy of the motion estimation. The methodology for the computation of the motion vectors is not specified by the standard and is left open as an design issue. There is of course a tradeoff between the accuracy of the motion estimation versus the complexity of the MC technique.

For **B** pictures, MC prediction and interpolation is performed using reference frames present on either side of it, where reference pictures include both **I** and **P** pictures. The prediction is non-causal, since it uses frames from the past and the future. Compared to **I** and **P**, **B** pictures provide the maximum compression. Some other advantages of **B** pictures include the reduction of noise due to the averaging process, and the use of future pictures for coding. This is particularly useful when coding"uncovered areas". **B** pictures are themselves never used for predictions and hence do not propagate errors. MPEG-1 allows for only frame based MC, while MPEG-2 allows for both frame and field-based MC. Field-based MC is specially useful when the video signal includes fast motion.

Pictures do not need to follow a static **IPB** pattern. Each individual picture can be of any type. However, for simplicity, a fixed **IPB** sequence is often used throughout the entire video stream. This is the procedure used in the study discussed in this chapter.

Image quality

Although MPEG-1 can be run at high bit rates and at full CCIR-601 resolution, it processes frames, not fields. That fact limits the attainable quality (even at data rates > 5.0 Mbps) and was responsible for motivating development of the MPEG-2 algorithm that can handle individual fields in the first place.

MPEG defines the syntax for storing and transmitting compressed data; decoding is fully defined. However, encoding is not defined! That fact notwithstanding, all conforming encoders must produce valid MPEG bit-streams that are de-compressible by any MPEG decoder. This is the key strength behind MPEG. By allowing encoders to effect proprietary, but compliant, algorithms, different mechanisms are plausible. If two encoders are each fed an identical video signal and equivalent output rates are maintained, there is no guarantee - nor even an expectation - that the compressed streams, or the de-compressed video quality, will be the same.

A robust MPEG compressor obtains its results, in part, by dynamically allocating bandwidth resources so that they are balanced among the details of each image. Some video features (an actor's facial nuances, for instance) are more valuable from the viewer's perception of quality than are others (such as the texture of the wall against which the actor is leaning). Encoding

details that do not add much (or even detract) to quality normally consume vast amounts of available bits that could otherwise be used to more faithfully represent the important attributes.

6.3 Experimental Methods

An assessment was made of the effects of MPEG-2 compression on the ability of human observers to recognise the class of a ship on viewing video sequences of the broadside view of vessels at sea.

6.3.1 Apparatus

The experimental setup consisted of a 486 66 Mhz DX personal computer (PC) with a Matrox PIP- 1024 image digitising and display card as the display driver and experiment controller. The image display device was an Electrohome 1719X high quality monochrome television monitor with a hood placed over the screen to constrain the viewing distance to 500 mm and to block out ambient light, though the room was darkened to maintain light adaptation. The video sequences were displayed via a PC MPEG-2 [9] board, which played the bit streams directly from disk running under the control of the experimental software that was developed for this purpose. There were many technical problems encountered in setting up the experimental system. These are outlined in Appendix F.

6.3.2 Procedure

During the experiment, ten observers each saw 512 five second long video sequences. These were presented in a different random order for each subject. The observer had to respond by pressing a computer keyboard key, then registering a class symbol and a confidence rating on a 5 point scale. Response time was also recorded. In addition to response time and confidence rating, the probability of recognition or hit-rate (as defined in equation 6.1), was used as a performance measure.

$$p_{r_i} = \frac{1}{N}\sum_j W_{ij}, \text{ where } W_{ij} = \begin{cases} 0 & \text{miss} \\ 1 & \text{hit} \end{cases} \text{ for the } i^{\text{th}} \text{ treatment and } j^{\text{th}} \text{ observer;} \quad (6.1)$$

where N = number of observers.

The experiment was performed in sessions with a maximum length of 30 minutes, to reduce fatigue or boredom effects. Prior to the main experiment, the observers underwent training to reduce practice effects.

Training

Prior to the main experiment, every observer underwent training sessions. This training was carried out using the experimental set up, but with uncompressed versions of the video sequences,

[9] Videoplex MPEG-2 board manufactured by Optibase.

to learn the ship classes and get used to the experimental procedure. Once they had achieved 100% hit-rate, they were considered suitable for the real experiment. This training was intended to reduce the learning affect, not associated with compression, during the experiment; *i.e.* to remove learning, not correlated with compression level, as a factor in the statistical analysis.

6.3.3 Stimuli

The original source of video sequences were taken from an airborne infrared sensor[10]. The experimental sequences were derived from 16 original sequences that contained 4 classes of naval vessel (see figure 6.2), which were compressed by different amounts and then de-compressed. Figure 6.3, shows frames for each class, which have been compressed and then de-compressed at 2.0 Megabits-per-second (Mbps).

6.3.4 Experimental Design

The experiment was a full factorial repeated measures design with fixed effects and blocking (Hines and Montgomery, 1980b), with the stimuli randomised within these blocks. The observers were treated as blocking variables, *i.e.* they were not included in the analysis and so the data was collapsed across them.

The fixed effect variables were the two discrete independent variables: target class (one of four ship class categories) and the compression ratio or bit stream rate (BSR) used on each sequence. The BSR was measured in Mbps and was restricted to the set 2.0, 4.0, 6.0, 8.0. Note an uncompressed video stream contains about 160 Mbps of information. This means in effect that the compression ratio set of 80, 40, 26.6, 20 to one was used.

Given that only 16 original sequences were available, the total number of compressed and de-compressed sequences generated by a full factorial design was equal to the number of original sequences times the number of compression levels times the number of target categories, *i.e.* $4 \times 4 \times 4 = 64$. This number of presentations was not sufficient to obtain a 95% confidence in the ANOVA analysis, which requires about 5000 presentations of stimuli spread over all the subjects.

As already stated, each subject observed 512 video sequences over the experiment. Ideally these should have been 512 separate compressed and de-compressed similar sequences of the targets under consideration. This would have reduced the possibility of observers learning particular video sequences, but would have been a logistical nightmare[11]. Therefore the additional stimuli were synthesised from the 16 available sequences by the following procedure.

The 64 original 704×576 frame-sized video sequences were compressed via MPEG-2 software[12]. During the experiment, under software control, the video sequences were de- compressed via the MPEG-2 decoder board and played on the video monitor. During a full experiment each video sequence was duplicated 8 times for each subject. In order to reduce learning affects,

[10] A UK Common Module IR imager (TICM II, 8-14(m) was mounted in the hatchway of an C-47 aircraft.
[11] See Appendix F for technico-logistical difficulties encountered.
[12] Each sequence took about 24 hrs to compress on a Sparc 10 workstation.

each 704×576 frame of video was windowed to 256×256 with the centre of the window being randomly placed within the larger frame. The placing of the window was constrained so that the target remained in view throughout the whole sequence.

Observer performance was assessed by analysis against the independent variables using Analysis-of-Variance (ANOVA) (Hines and Montgomery, 1980a).

6.3.5 Informal Study on Temporal Processing Gains

A quantitative study to determine the gains in target acquisition through the temporal dimension, was considered to be useful, but was not undertaken due to time and resource constraints. However, I undertook an informal study comparing the detectability of still and moving target sequences, and presented the same stimuli to 5 of the subjects, after the main experiment. The data was only qualitative, but gave some insight on the gain in target detectability achieved by the HVS through temporal processing.

6.3.6 Degradation to Failure

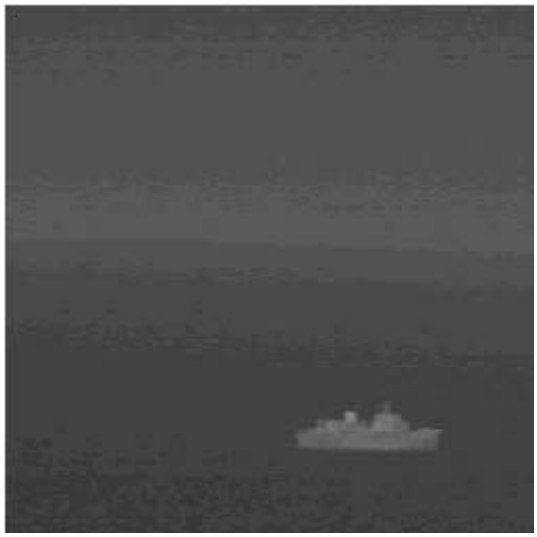
Good experimental design requires that the stimuli cover the full perceptual range: the compression-induced degradations must produce a range of effects so that target recognition ranges from extremely difficult all the way to relatively easy. Certainly degradation must be sufficiently bad that some observers will fail to recognise some sequences.

However when running the MPEG-2 encoding software with the parameters in default mode, it turned out to be impossible to produce bit streams below 2.0 Mbps without getting buffer overflow errors. At this level of compression there was not enough apparent degradation in the video. Therefore, it was deemed necessary to first degrade the video before compression. This was first done by adding white noise and lowering the contrast. Although, the original video (from an infrared sensor) was of high quality, real surveillance imagery can be severely degraded due to atmospheric and other effects, so this degrading process was acceptable. Since this still did not produce absolute failure in the recognition task, the next approach taken was to artificially double the field of view (*i.e.* to halve the angle subtended by the target), and then degrade. Pilot studies indicated that failure to recognise was achieved, at least for the lower bit rates.

(a) River class frigate.

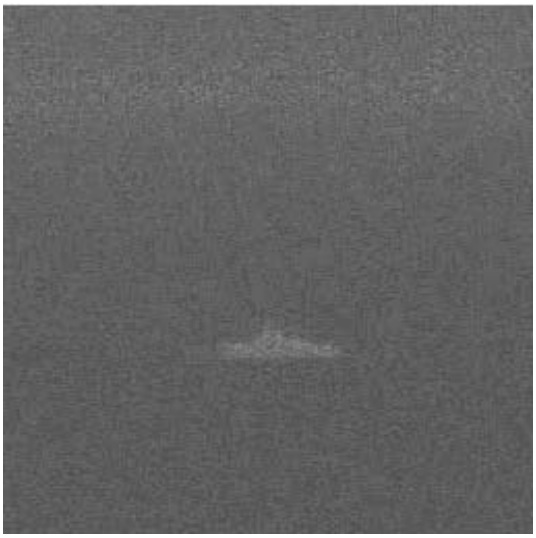
(b) Patrol Boat

(c) Perth class frigate.



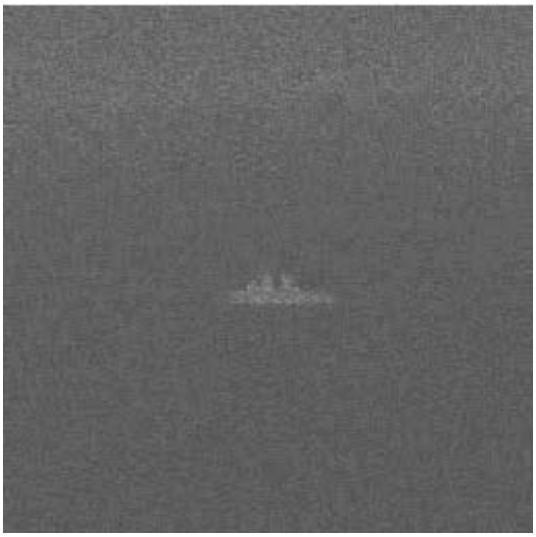
(d) Destroyer Tender.

Figure 6.2: Examples of $18.5^\circ \times 18.5^\circ$ (256×256 pixel) regions from from original video frames.

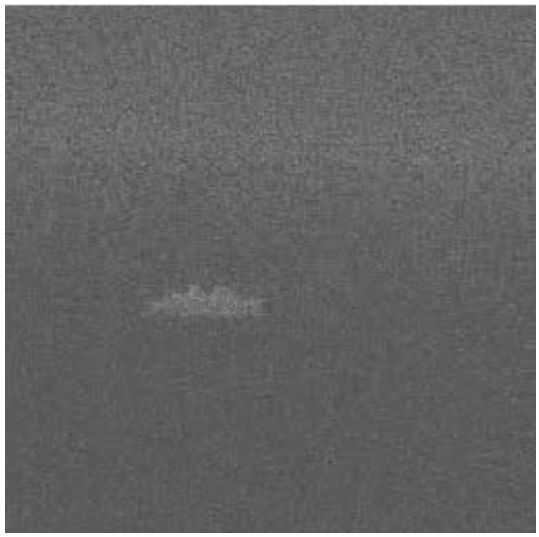


(a) River class frigate at 2.0 Mbps.

(b) Patrol Boat at 2.0 Mbps.



(c) Perth class frigate at 2.0 Mbps.



(d) Destroyer Tender at 2.0 Mbps.

Figure 6.3: Examples of $18.5^\circ \times 18.5^\circ$ (256×256 pixel) regions from de-compressed frames.

6.4 Results

Figure 6.4 shows a graphical summary of the results of the experiment, excepting trend analysis, which is shown in figures 6.5 and 6.6 in section 6.4.4. The probability of acquisition at different compression ratios varied only over a small range (96% - 100%, as shown in figure 6.4(a)) and, as expected, showed no statistical significance ($p > 0.69$) for any class, which indicates that hit-rate was not useful as a measure of target acquisition in this experiment; *i.e.* hit-rate was too insensitive. However, the near constant high level of the hit-rate suggests that *response time* was a valid measure of observer performance (for reasons discussed in section 5.3 in Chapter 5), and it showed as statistically significant the effects on observer performance.

The confidence rating was not used in this analysis. This was collected to develop some form of recognition receiver operating characteristic, but it was decided not to proceed with this approach.

6.4.1 Interactions

Figure 6.4(d) shows the interaction of class with compression level, which was shown by the ANOVA to be statistically significant. However, perusal of figure 6.4(d) indicates that this can mainly be accounted for by the interactions between class at values of 'patrol' and 'perth' at compression levels 2 MB/s and 8 MB/s. This may explain the apparent change in the trend of decreasing response time with increase in compression level (this is discussed in section 6.4.2), as shown in figure 6.4(b), for a compression level of 8 MB/s. This will not effect the main observation gained from the inspection of figure 6.4(c), which is that the latency in response time is longer for the river class ship target than any of the other classes.

6.4.2 Performance versus Compression Level

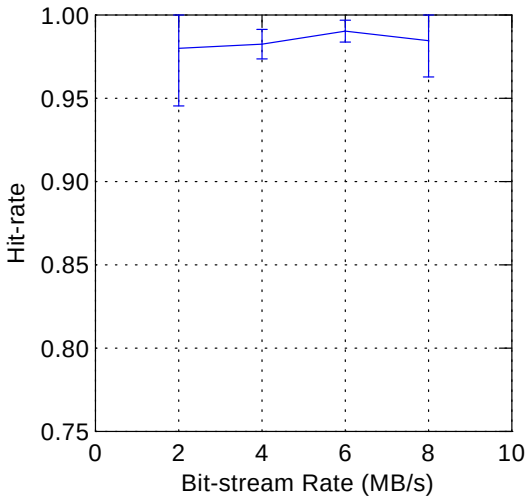
Figure 6.4(b) shows the plot of mean response time versus compression level in Megabits per second (Mbps). The observer performance at the 2.0 Mbps level was significantly slower than at the other levels of compression. However, the post-hoc analysis, which is shown in table 6.1, revealed that the other apparent differences in performance at the compression levels of 4.0 Mbps - 8.0 Mbps were not statistically significant; *i.e.* the other differences shown here most likely occurred by chance and are not due to any real affect.

6.4.3 Performance versus Target Class

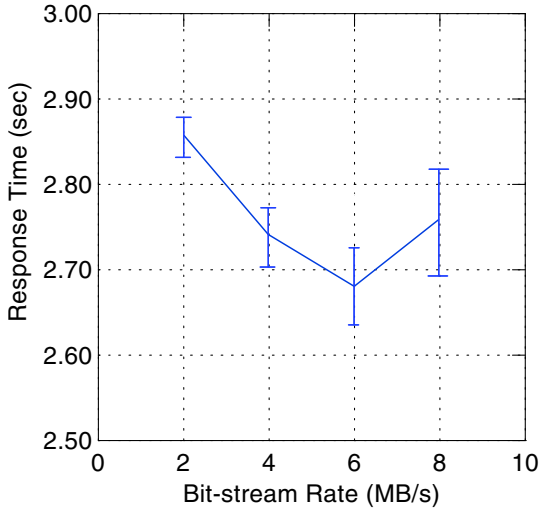
Figure 6.4(c) depicts the main effect for the independent variable of target class with response time as the dependent variable. It was found that there was an overall statistically significant effect of ship class upon observer performance. Post-hoc analysis (shown in table 6.2) indicated that this effect was predominantly due to the observer having more difficulty in distinguishing the "River" class frigate from the other ship classes. This finding agrees with the reports of the observers, gained during informal debriefings after completing the experiment. Everybody

Matrix of pairwise comparison probabilities				
Bitstream Rate	2.0	4.0	6.0	8.0
2.0	1.0000			
4.0	0.0038	1.0000		
6.0	0.0001	0.9615	1.0000	
8.0	0.0150	0.9863	0.4306	1.0000

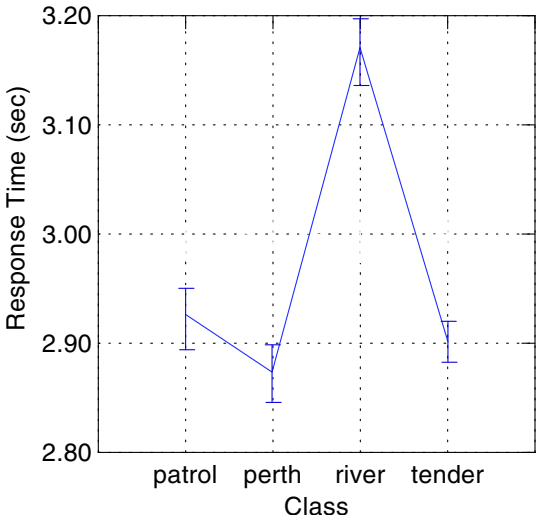
Table 6.1: Table shows post-hoc comparisons (Bonferroni) of response time for the 4 compression levels (MB/s), where the elements of the table are p-values.



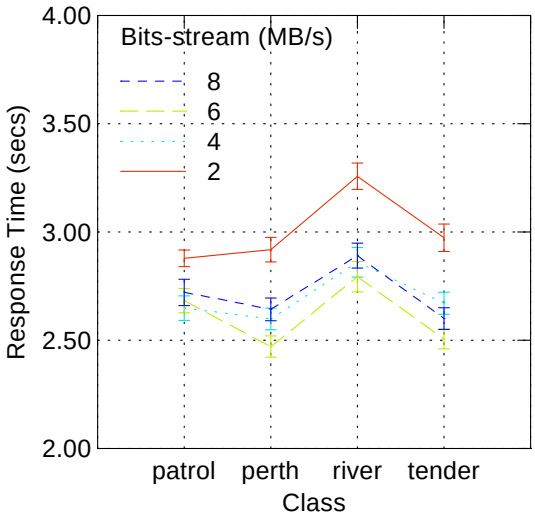
(a) The main effect for compression level, with hit-rate as dependent variable.



(b) The main effect for compression level, with response time as dependent variable.



(c) The main effect for class, with response time dependent variable.



(d) The interaction of class with compression level.

Figure 6.4: Graphical summary of effects. Compression level is specified in megabits per second (MB/s), response time in seconds, and hit-rate is dimensionless. Error bars denote standard deviations.

commented that they had most difficulty discriminating between the River class and one of the other class of ships.

Matrix of pairwise comparison probabilities				
Bitstream Rate	patrol	perth	river	tender
patrol	1.0000			
perth	1.0000	1.0000		
river	0.0000	0.0000	1.0000	
tender	1.0000	1.0000	0.0000	1.0000

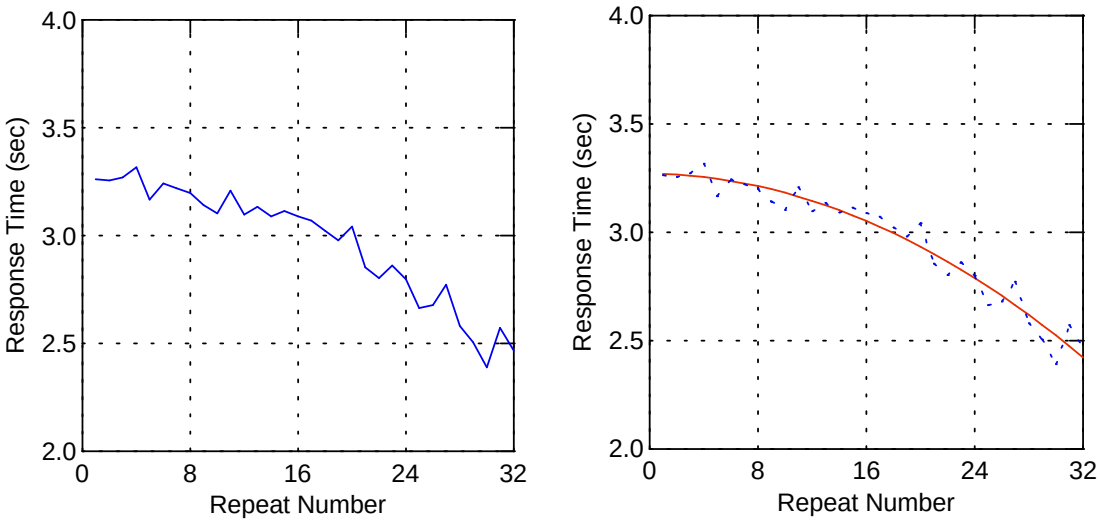
Table 6.2: Table shows post-hoc comparisons (Bonferroni) of response time for the 4 target classes, where the elements of the table are p-values.

6.4.4 Learning Effects

Despite the training regime that all subjects underwent prior to the experiment, it was found that there was still a significant learning effect evidenced. This can be seen in figure 6.5(a), which is a graph of response time (t_r) versus the repeated presentation (r_p) of the same stimulus set[13]. It is obvious that the non-linear trend in the graph is a progressive reduction in response time; *i.e.* learning is taking place. A regression analysis shows this curve to be a quadratic ($t_r = 3.2707 - 0.0009r_p - 0.0008r_p^2$), with an extremely good fit ($R^2 = 0.993$), as shown in figure 6.5(b).This seems to go against conventional wisdom, which espouses the *power law of practice* (Newell, 1990) for general learning affects. However, in normal practice, reaction time is plotted against trial number, and using logarithmic axes. This procedure was performed, and the results are shown in figure 6.6, which indicates that the power law of practice is approximately obeyed. These data represent the mean response time over subjects in order of presentation of stimuli. Note, the order of stimuli for each subject was randomised. Therefore, the actual stimuli seen by different subjects at a particular trial were probably different. Usually, in experiments determining the power law of practice, the difficulty of each trial has been about the same. However, here the difficulty of each stimulus was not equal (as borne out by the ANOVA). This may explain some deviation from the expected power law of practice.

The full within-subjects (repeated measures) ANOVA table for this experiment is shown in table 6.3.

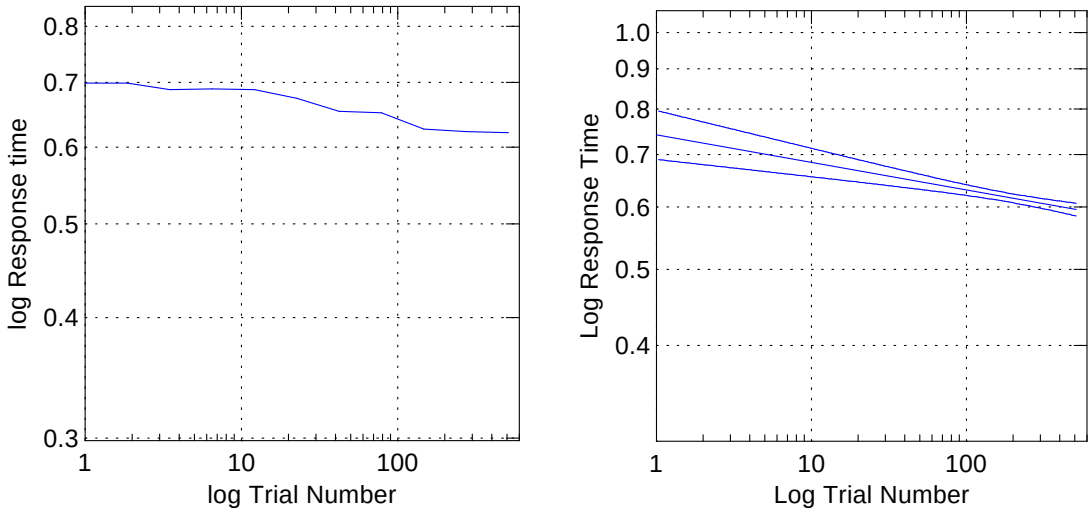
[13]That is the full set of video sequences of the 4 ship classes each at the 4 compression levels.



(a) Mean response time for repeats of 4×4 factorial experiment.

(b) A quadratic regression fit to learning curve ($R^2 = 0.993$).

Figure 6.5: Learning effect for repeats of complete stimulus set.



(a) Log-Log plot of t_r versus trial number.

(b) A linear regression fit (for log axes) learning curve ($R^2 = 0.987$).

Figure 6.6: Learning effect over individual stimulus trials.

Summary of all Effects for ANOVA					
Effect	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	p-level
sub stratum	9	42.3879	4.7098	-	-
sub.trial stratum					
trial	31	39.3771	1.2702	2.14	<.001
Lin	1	1.5547	1.5547	2.61	0.107
Quad	1	20.1546	20.1546	33.88	<.001
Cub	1	0.0609	0.0609	0.10	0.749
Deviations	28	17.6068	0.6288	1.06	0.391
Residual	279	165.9520	0.5948	-	-
sub.compress stratum					
compress	3	48.6398	16.2133	46.28	<.001
Residual	27	9.4585	0.3503	-	-
sub.class stratum					
class	3	1831.7893	610.5964	460.37	<.001
Residual	27	35.8102	1.3263	-	-
sub.trial.compress stratum					
trial.compress	93	128.9504	1.3866	3.24	<.001
Lin.compress	3	6.9096	2.3032	5.38	0.001
Quad.compress	3	39.6247	13.2082	30.85	<.001
Cub.compress	3	16.7426	5.5809	13.04	<.001
Deviations	84	65.6735	0.7818	1.83	<.001
Residual	837	358.3451	0.4281	1.19	-
sub.trial.class stratum					
trial.class	93	91.8116	0.9872	2.60	<.001
Lin.class	3	14.4259	4.8086	12.66	<.001
Quad.class	3	19.4425	6.4808	17.07	<.001
Cub.class	3	5.6424	1.8808	4.95	0.002
Deviations	84	52.3008	0.6226	1.64	<.001
Residual	837	317.7938	0.3797	1.06	-
sub.compress.class stratum					
compress.class	9	451.1267	50.1252	82.96	<.001
Residual	81	48.9410	0.6042	1.68	-
sub.trial.compress.class stratum					
trial.compress.class	279	853.3746	3.0587	8.53	<.001
Lin.compress.class	9	476.6599	52.9622	147.64	<.001
Quad.compress.class	9	142.4159	15.8240	44.11	<.001
Deviations	261	234.2988	0.8977	2.50	<.001
Residual	2511	900.7607	0.3587	-	-
Total	5119	5324.5185	-	-	-

Table 6.3: ANOVA table with response time as the dependent variable.

6.5 Discussion and Conclusions

The data from this study demonstrate an obvious effect of video compression on observer performance in target recognition. However, this effect was not large enough to cause serious misclassification of targets and was largely due to the degradation caused at the maximum video compression (2.0 Mbps). The degradation caused by compression did increase the time required for target recognition, particularly at the highest compression level. In a real world setting, depending on the application, this may introduce an intolerable degradation in observer performance.

Ship classes were chosen so as to introduce a range of difficulties in discriminability between target classes. However, this was limited by the range of suitable video footage available, so that four classes was considered a minimum sized set. Under these constraints, the set of four ship classes, as described earlier, was chosen subjectively by the author. After performing the experiment, it became evident that similarity or otherwise of targets was very subjective. This information was mainly gleaned from the post-experimental debriefing of the observers, but it was also evident in the data. In either case, from both anecdotal and statistical evidence, all observers found the River class vessel the hardest to discriminate from each of the other vessels. This suggests that the features of the River class ship overlap with the features of the other three classes in perceptual space, and further exploration of this would require techniques such as Multi-Dimensional Scaling (Evans and Attaya, 1978).

There was a significant learning effect, even after the initial training sessions. In other words, the observer's performance kept improving with practice. This means that the human visual system (HVS) was learning the distortions introduced by compression and either compensating for them or modifying the internal model of the target. This has serious implications for training of personnel using systems which require compressed imagery, as it shows that operator efficiency can be greatly improved by training in performing surveillance related tasks. Of course, there would be a point after which further training would give only minimal improvement in performance. This point was not indicated in the study described here.

As described in section 6.3, the upper limit for compression, allowed by the MPEG-2 software, was at 2.0 Mbps (approximately 80:1), and this level accounted for most of the degradation. There is an emerging standard for low bit rate video compression called MPEG-4, which would allow much lower bit rates to be obtained. This standard is being designed for many applications including surveillance systems. The experiments should have been based on, or at least included, MPEG-4 compression. However, at the time this experiment was carried out, there were no MPEG-4 software or hardware encoding systems available.

Some of the sequences produced were extremely degraded after compression and with added noise and contrast reduction (prior to compression). As a result, individual still frames were quite difficult if not impossible to interpret on their own. This suggests that the temporal processing of the human visual system is a powerful aid in detection and recognition, and provides further motivation for pursuing this research. To illustrate this, figure 6.2 shows typical frames from de-compressed sequences. Although obviously degraded, temporal integration still allows an

observer to recognise the ship from the associated sequence.

6.5.1 Implications for Task Related Video Quality Metrics

This work has some general implications for video quality metrics[14], and has particular relevance to video compression. In such applications, the temporal dimension has to be considered first. Even though this aspect of the work was only touched on, it was clear that temporal processing in the HVS made a significant difference to the ability of observers to classify targets. Therefore, any video quality metric must be applied over several video frames, and be applied both spatially and temporally. How many frames need be "sampled" and the relative weighting of spatial and temporal properties to be included in a metric, are factors which are probably dependent on the visual task and application, and, as such, are subjects for further research.

In considering a metric which will predict human performance, cognisance has to be taken of the learning effect demonstrated here. Since vastly different performances can be achieved for the same video stimuli at different points in the learning curve, this means that an appropriate metric will allow for the type of visual task and the level of observer experience. This may mean that basic metrics work under a "meta" metric. (Such a system is discussed in the final chapter.)

This study has shown that when dealing with higher level visual tasks, such as target recognition, as against simple detection, variability between individuals increases. However, though individual differences in subjective response were apparent, it was noted that, every subject found one particular class of target equally the most difficult to classify. These commonalities and differences in individual subjective perception, were discussed in section 6.4.3 and may impact on the design and application of an image metric, when applied to higher level visual tasks. As an example consider a metric, that is required to measure the degree of difficulty a human would encounter in classifying targets in various video sequences. This metric would attempt to measure the degree of separation of targets in subjective perceptual space. It may achieve this by transformimg appropriate physical video image characteristics, by means of a suitable mathematical construct, into a map of the target's features in a space, analogous to the human perceptual space. However, there appears to be some variability in how humans map the target features into perceptual space and these need to be considered in the metric design.

[14]The term "metric" is not being applied here in the strict sense defined in Chapter 2.

Chapter 7

Effects of "Local" Clutter on Human Target Detection

Summary: *This chapter deals with the effects of clutter on human target acquisition. In theory, properties of clutter can be defined globally or locally. However, in the literature, the distinction between local and global clutter is arbitrary. If the image contains different clutter types, global clutter metrics may be inappropriate and are expensive to compute. It is also likely that, in terms of detection, rather than search, local clutter is more important. The problem is to determine how local is local? In the literature, the standard approach of setting the local domain to twice the expected target size is adopted without any justification. This chapter addresses this problem and considers the implications for the application of clutter metrics.*

7.1 Introduction

This chapter deals with the effects of the extent of the clutter around a target on the human's ability to detect that target. Here, clutter is defined as any structure in the image, which masks the target or confuses the observer as to the location and/or class of the target. In theory, properties of clutter can be defined globally or locally (Rotman et al., 1994), and it has been shown that the HVS processes information at both the global and local levels (Caelli and Julesz, 1979; Burr et al., 1979). Whether or not the HVS uses mainly local or global pre-attentive[1] cues to focus attention is determined by the properties of the particular image being viewed. Apparently, the HVS can integrate global features, such as texture statistical properties (Caelli and Julesz, 1979), or underlying spatial spectral distributions (Burr et al., 1979), which in turn tune the HVS to local properties.

These interactions between global and local properties in visual processing must depend on the clutter and target types in the detection context. This study cannot claim to explore all these possibilities, but is intended as a preliminary investigation of the effect of clutter localisation on human target detection and of its implications for the application of local clutter metrics.

[1] See Chapter 2 section 2.6.1.

This is important, since it is conventional wisdom to set the local domain to twice the expected target size, without any justification. It is also important because it is likely that, in terms of detection, rather than search, local clutter has a greater effect than global clutter on visual performance (Overington, 1976b; Doll et al., 1993).

The research described in this chapter is expected to facilitate the following aims by providing some information on the extent and functional form of the effects of local clutter on human target detection.

- To give more accurate prediction of human target detection, by using clutter metrics which will be more representative of human visual response;
- To increase the efficiency of the computation of clutter metrics, as images with near homogeneous clutter level were expected to require only local extent to be computed. Other images were expected to require computation for only a few instances in regions of homogeneous clutter level;
- To gain a better understanding of human vision.

7.2 Experimental Methods

The stimuli presented to the experimental subjects consisted of circular regions of simulated background clutter, at the centre of which existed a circular region (target), with an incremental increase in luminance[2] (ΔL_t) over the rest of the clutter region. The surround consisted of a uniform luminance (L_s), equal to the average simulated clutter luminance (μ_b).

For each given background luminance and target radius (r_t), the relationships were explored between background disc radius (r_{clut}), target to background contrast (c), a single parameter (δ) controlling the clutter statistics and the probability of detection (p_d) of the target.

Circular targets and clutter background regions were used in this study because many studies in the visual detection literature have used circular stimuli. However, previous studies have used uniform luminance targets and/or uniform luminance backgrounds/surrounds. A classic example is that of Blackwell (1946), who defined human visual contrast thresholds using uniform disc targets on uniform backgrounds. Most of these studies do not address the interaction between size of the background and visual performance, although Overington does cite some studies in his book (Overington, 1976a), which addressed this issue with uniform luminance stimuli. On the basis of these studies, he stated that size of the surround has little effect if this is less than 1 to 2 orders of magnitude less bright than the target (Overington, 1976c) (positive target contrast) and the local surround is greater than about 6 milli-radians ($\approx 0.3^\circ$). However, if the surround is much brighter than the target (negative target contrast), the detection threshold drops markedly. For local target surrounds of a size range of about 1 to 2 milli-radians, a decrease in detection threshold has been reported (Overington, 1976d). These results provided

[2]Each pixel in the target region had a constant luminance increment applied.

a base-line for this study, and show that any effects related to the (clutter) background size are related to the clutter effects and not purely luminance-surround size interaction effects.

Now that a basic description of the experiment has been given, the experimental design and other details are explained in the remainder of this section.

7.2.1 Experimental Design

The experiment was a full factorial, fixed effects repeated measures (within-subjects) design. There were four factors in the design, namely - clutter property δ (see section 7.2.2), clutter background radius r_{clut}, target radius r_t and target contrast c.

There were 2304 treatments, presented to each subject in a different random order. These consisted of all the combinations of the four factor levels; *i.e.* 4 levels of $\delta\times$ 12 levels of $r_{clut}\times$ 6 levels of $r_t\times$ 4 levels of $c = 1152$. There was also an implicit fifth factor, target presence with levels 0 or 1; *i.e.* for each stimulus containing a target there was a stimulus with identical clutter properties and background radius, but no target. Thus total number of stimuli was $1152 \times 2 = 2304$.

7.2.2 Image Stimuli

As explained in the previous section, each of the 2304 treatments presented to each subject represented a different combination of the levels of the four factors δ, r_{clut}, r_t, c, with and without a target. Obviously $r_t = c = 0$ for no target in the image. Each of these treatments corresponds to a unique visual stimulus when a target is present. For the stimuli without a target present there is uniqueness only in the factor combinations of clutter radius and clutter property (δ).

As previously, the (Weber) contrast was defined as

$$c = \frac{\Delta L_t}{\mu_\mathrm{b}} = \frac{\mu_\mathrm{t} - \mu_\mathrm{b}}{\mu_\mathrm{b}} \tag{7.1}$$

where μ_t is the mean target luminance and μ_b is the mean local background luminance. For reasons given in section 7.2, only positive target contrasts were considered in this experiment.

Figure 7.1 shows some examples of the visual stimuli presented to the subjects, for the different clutter types. As mentioned earlier, these stimuli consisted of circular regions of simulated background clutter, with variable radius, co-centred with a circular region of smaller radius, generated by an incremental increase in luminance. The surround with luminance equal to the average simulated clutter luminance. Table 7.1 gives the values of the stimuli variables that were used in the experiment.

The target to background luminance contrast was calculated "on the fly" during each trial. This was done by using the screen grey-level to luminance data that was measured prior to the experiment, and is represented by the the regression formula in equation 7.4. However, it was found that using equation 7.4 to determine luminances sometimes resulted in large errors. This was overcome by linearly interpolating between the actually measured points on the video

Stimulus Factors and Levels			
Factors			
Background Radius	Target Radius	Contrast	Clutter Parameter (δ)
Levels			
1.4	0.3	2.0	-0.050
2.1	0.5	3.3	-0.300
2.8	0.7	4.7	-0.550
3.5	0.9	6.0	-0.925
4.2	1.1		
4.9	1.3		
5.6			
6.4			
7.0			
7.7			
8.4			
9.2			

Table 7.1: Table shows level values for the independent factors, which are the stimulus variables. The radii are given in degrees.

monitor luminance calibration curve, which bounded the error on the screen luminance contrast presented to the subjects.

The images as shown in the figures are not photometrically correct, as they indicate grey-levels and not luminance. These clutter types are represented by the parameter δ, which in effect determines the granularity or "clumpiness" of the images.

Simulated Clutter

A compromise was required in setting up the visual stimuli. On the one hand, the background clutter needed to represent real clutter. On the other, there was a need to have control over the image statistics. As it happened, colleagues were working on simulating natural vegetation from a knowledge of the statistics of natural imagery. They were doing this by generating simulated images with the same statistics as their real image counterparts, in particular by ensuring that the simulated and real images had a similar auto-covariance function (Bertilone et al., 1997; Bertilone et al., 1998). Bertilone and colleagues showed that imagery in the visible spectrum have a Gaussian distribution of grey-levels over an ensemble of images. However, in order to control the statistics of the simulated image, several parameters were required to be manipulated. In light of some work done by my supervisor (Newsam and Woodruff, 1991), who suggested that fractal image stimuli would simplify experimental design and setup (this is fully discussed in section 7.2.4 on page 136), I thought it appropriate to consider a modification to the work of my colleagues, to produce a simulation of clutter imagery which had random Gaussian statistics but was fractal in nature.

The background clutter in the stimuli were derived from a fractal statistical process. This

was done for two major reasons:

- Natural scenes can be described well with a fractal model (Pentland, 1984; Knill et al., 1990; Rotman et al., 1994; van der Schaaf and van Hateran, 1996);
- To allow control of the process by a single parameter. This was to minimise the size of the experimental design;
- To simplify the physical design of the experimental setup[3].

Figure 7.2 shows examples of images with the four values of δ used in the experiment.

Fractal Simulated Clutter Algorithm

The following is a brief background to the generation of the fractal image stimuli and the meaning of the δ parameter. This is also an introduction to the full mathematical derivation for the fractal image generation algorithm which is given in Appendix G.

Consider an image, from a family consisting of N images (an *ensemble*), which are *isotropic*[4], *stationary*[5], *Gaussian random fields* (GRF) (Yaglom, 1987), with $L_i(\tilde{x})$ denoting the luminance function of the i^{th} image (*i.e.* $i \in \{1, 2, \ldots, N\}$) in the spatial region $R \equiv \{\tilde{x} = (x, y) : 0 \leq x \leq X, 0 \leq y \leq Y\}$. This family of images will form a *realisation* of a Gaussian random field if, the N sized set of luminance values $L_i(\tilde{x})$ for each point $\tilde{x}$ in the field, is a sample from a Gaussian distribution with mean $\mu(\tilde{x})$ and variance $\sigma^2(\tilde{x})$. With this condition met, every pair of luminance values $L_i(\tilde{x})$ and $L_i(\tilde{y})$ at two different spatial locations $\tilde{x}$ and $\tilde{y}$, form a bivariate Gaussian distribution[6]. A Gaussian random field image $L(\tilde{x})$, is completely characterised by its mean and its covariance function $C(\tilde{x}, \tilde{y})$, which is defined as

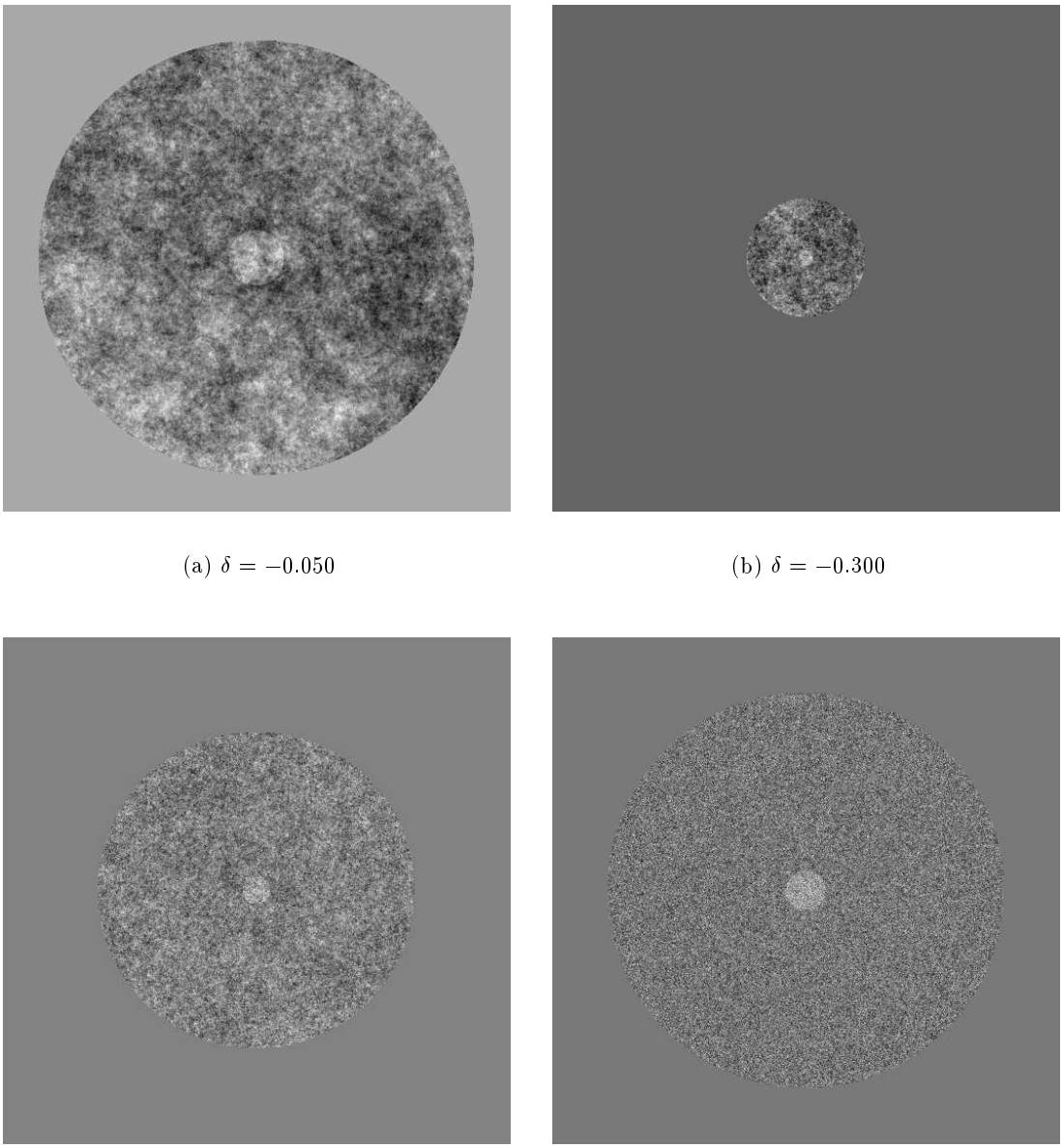
$$C(\tilde{x}, \tilde{y}) = \langle (L(\tilde{x}) - \mu(\tilde{x}))(L(\tilde{y}) - \mu(\tilde{y})) \rangle \,, \tag{7.2}$$

where $< \cdot >$ is the expectation operator. Since the GRF is isotropic and stationary, $r = ||\tilde{x} - \tilde{y}||$; *i.e.* $C(\tilde{x}, \tilde{y}) = C(r)$. For the GRF to be fractal, *i.e.* invariant under scaling, it is required that

$$C(r) = kr^{2\delta}, \tag{7.3}$$

where $\delta = \frac{\ln t}{\ln s}$, with $s \neq 1$, a scaling factor on the region size, and t a scaling factor on the luminance values within the region (Newsam and Woodruff, 1991). Note, δ must lie in the range $-1 < \delta < 0$, and is the single parameter which controls the "roughness" of the texture of the fractal image (as shown in figure 7.2). The fractal GRF thus defined, which has $\mu = 0$ and infinite variance, can only exist notionally, but can be realised by convolving (*i.e.* smoothing) it with some sensor function (*e.g.*, the eye).

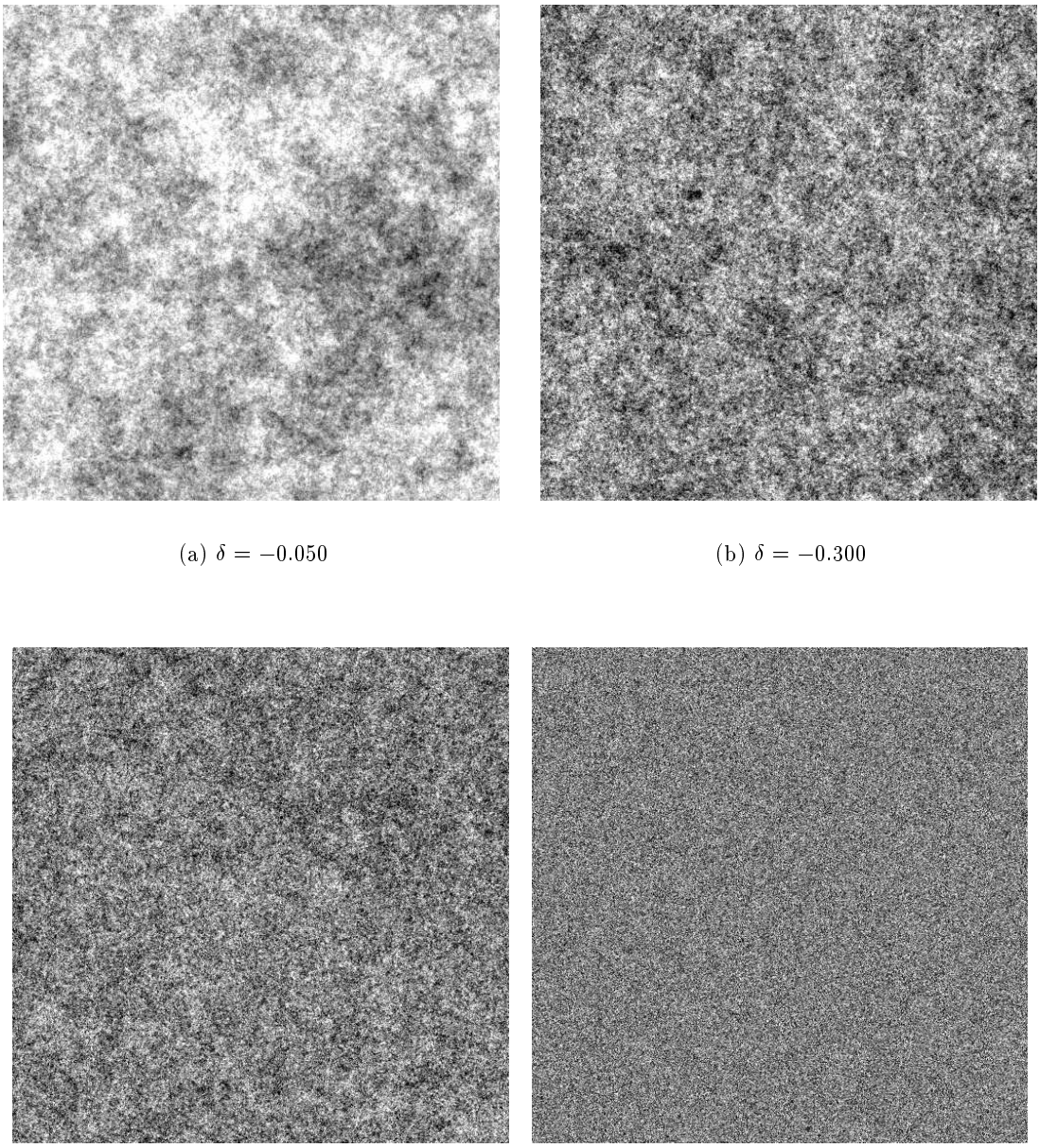
[3] This proved not to be as simple as first thought, see section 7.2.4.
[4] The statistics are independent of direction;*i.e.* invariant under rotation.
[5] The statistics are the same at all regions within the image; *i.e.* invariant under translation.
[6] Where $\tilde{y} = (x, y), \neq \tilde{x}$.



(a) $\delta = -0.050$ (b) $\delta = -0.300$

(c) $\delta = -0.550$ (d) $\delta = -0.925$

Figure 7.1: Illustrations of stimuli seen by the experimental subjects. The clutter properties are controlled by δ the clutter parameter.



(a) $\delta = -0.050$ (b) $\delta = -0.300$

(c) $\delta = -0.550$ (d) $\delta = -0.925$

Figure 7.2: Background clutter images for different values of clutter parameter δ.

Pilot Study on Simulated Clutter Perceptual Scaling

As discussed above, the clutter was controlled by a single parameter, δ. In order to test how this parameter is related to the subjective perception of clutter properties, a pilot study was carried out prior to the main experiment. A series of hard-copy images was produced, similar to those shown in figure 7.2, with δ incremented from -0.5 to -0.95, in steps of -0.5. These were presented to seven subjects[7], who were asked to order the images and then place them on a bench, such that the physical distance between the images represented the perceived distance between the "clumpiness" or roughness of their texture. Each subject participated in isolation from the others, and the image stimuli set was newly shuffled for each subject.

Of the seven subjects, two failed to completely order the images perfectly with respect to δ. However, even these two subjects ordered the images in a manner that was correct overall, with "mistakes" taking the form of the reversal of pairs of adjacent images (with respect to δ). Figure 7.3 depicts the scatter-plot of the normalised mean subjective rating (measured distances between photographs) versus δ, with the linear regression line overlaid. It is evident from figure 7.3 that the parameter δ corresponds to a subjective linear mapping of the simulated clutter property of clumpiness. This then indicated that δ was appropriate as a factor in the experimental design, particularly for analysis-of-variance, which is based on a linear model.

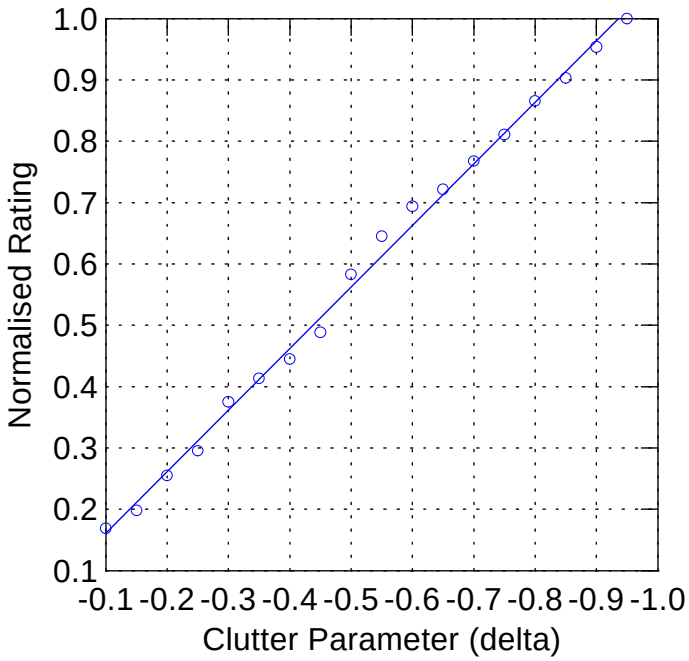


Figure 7.3: The subjective rating of clutter "clumpiness" ($R^2 = 0.996$).

7.2.3 Experimental Procedure

The subjects were all volunteers from within the age group 28 - 50 years, and had at least 6/6 vision. They were instructed as to how to conduct the experiment by the following means.

[7] Only two of these subjects participated in the main experiment.

- Written instructions detailing the use of the software and the experimental procedure. (See Appendix H).
- Verbal instructions, with the same information and an opportunity for clarifying any unclear points.
- A demonstration and trial period with a training version of the software that gave feedback, after each trial, on target presence and size.

The software presented the stimuli and logged response times in conjunction with the observer's confidence rating for each trial. The observers were required to find targets which had been placed into imagery according to the principles discussed in chapter 3. Observers were then prompted to enter their confidence rating according to a 5-point scale.

Each session was arranged for the same time every day for each observer and was limited to a maximum duration of one half-hour. After an initial training session, each observer typically sat through 4-6 experimental sessions. The complete instructions given to the subjects are given in Appendix H.

7.2.4 Apparatus

A 486 66 MHz DX personal computer, which controlled a video monitor, that had its photometric output to grey-level input recorded, ran experimental software especially written (in Borland C) to control the experiment. Images were presented on an Electrohome 1719X high quality monochrome television monitor from a Matrox PIP-1024 image digitising and display card. A hood was placed over the screen to constrain the viewing distance to 500 mm and block out ambient light, which was held to a constant low level as the experiment was performed in a light controlled room.

Implications of Fractal Images on Setup

As outlined in section 7.2.2, there were two main reasons for selecting fractal images for the stimuli in this study. Firstly, with large factorial experiments, it is easy for the size of the experiment to become unwieldy. In order to reduce the number of variables in this study, the control of clutter properties by a single parameter, as shown in equation 7.3 of section 7.2.2, was attractive.

Secondly, Newsam and Woodruff (1991) argue that using fractal GRFs as stimuli rather than standard GRFs, simplifies the experimental setup. They present a hypothetical scenario, where an ideal[8] observer is able to perceive directly a fractal GRF which is characterised by equation (7.3). They argued that, in this case, the perceived image statistics would be independent of the viewing distance and thus pixel size; *i.e.* the perceived image would be scale invariant. In contrast, they showed that for a standard (non-fractal) GRF image, the perceived

[8]Ideal in the sense, that the observer's vision does not undergo blurring; *i.e.* the eye PSF is a δ function.

correlation between the pixels depended on the perceived pixel size; *i.e.* the image statistics changed with viewing distance.

At the same time, as Newsam and Woodruff pointed out, a GRF characterised by equation (7.3) cannot be physically realised. Even if it could, once it is converted to a digital image format, its properties change. In fact, the algorithm (see Appendix G), used to generate the fractal clutter image here, produces each pixel of a digital image by integrating the conceptual fractal GRF over a defined region. It is to this region which becomes a pixel in the produced digital image, that Newsam and Woodruff's arguments apply, in the sense that, whatever the size of this region of integration, the resultant image statistics will remain constant. This is in contrast to a continuous, standard GRF, which, if digitised by integrating over regions of specified size to obtain pixels, would have statistics in the digital image that depend on the scale of the region of integration.

Once the 'fractal' digital image has been displayed on a monitor, with its fixed pixel size, a correlation length is imposed and the image is no longer scale invariant; *i.e.* a correlation function is defined. The effect of the parameter δ on the correlation function is shown graphically in figure 7.4. As the viewing distance increases, the angle subtended by a display pixel on the retina of the viewer decreases and thus the angle subtended by a distance equivalent on screen to the correlation length decreases. The perceived correlation between the pixels will decrease as the ratio of the angle subtended by the correlation length and the angular separation of the photoreceptors in the retina falls. As this ratio falls below unity, the image will rapidly look like white noise.

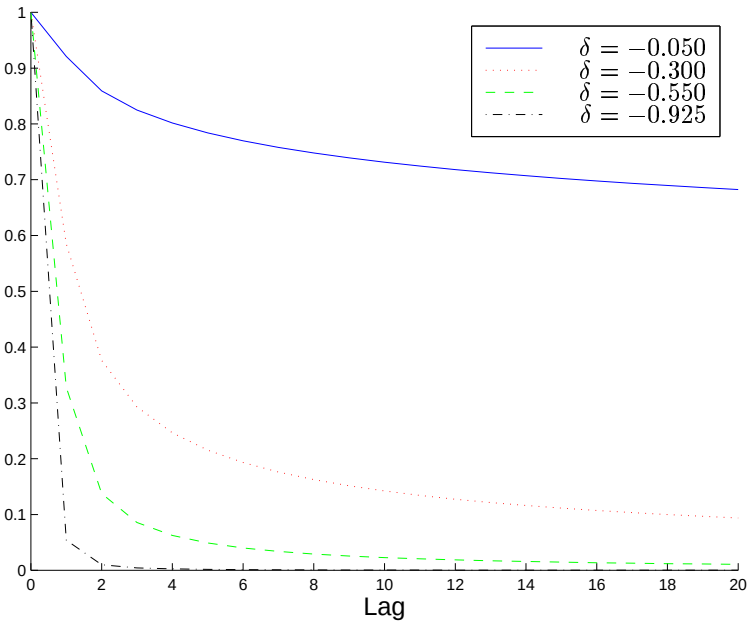


Figure 7.4: The effect of the clutter parameter (δ) on inter-pixel correlation. This figure shows the autocorrelation function for each of the four values of δ. The abscissa indicates the lag in pixels, while the ordinate is the amplitude of the autocorrelation function, which has dimensionless values between 0 and 1.

Therefore, I believe that the assertion of Newsam and Woodruff that a fractal GRF image, when used as a stimulus in a visual experiment is scale invariant, is incorrect, even though their mathematical development appears correct; *i.e.* their argument is correct. However, it applies at the stage where the realisation of a digital image is formed (by the fractal image generation algorithm), not at the display of that image.

The reasons for using fractal images as stimuli were discussed in section 7.2.2. The following advantages were expected to be gained from using fractal imagery:

(i) A realistic simulation of natural scenes;

(ii) The control of image statistics by a single parameter; and

(iii) The simplification of the experimental setup, due to the insenstivity of perceived image statistics to viewing distance.

From the point of view of this chapter, the advantages listed in points (i) and (ii) are the most important, and were gained by use of the fractal image generation algorithm. However, the expected advantage mentioned in point (iii) was not achieved for reasons discussed above. This became inconsequential, since the experimental apparatus included a viewing hood to fix the viewing geometry of the observers.

Measuring screen luminance

The screen luminance for each possible grey-level was measured using a photometer. This ensured that the results could be meaningfully compared with others in the literature. To achieve this, a full screen for each grey-level was displayed and a luminance measurement taken, with the photometer focused at the centre of the screen. Samples were also taken in each of the four corners of the screen to test for uniformity in luminance across the screen. The results of these screen luminance measurements were used in equation (7.1) to calculate the target contrast. The relation between screen luminance, L, in cd/m^2 and pixel grey-level value, g, was closely approximated by the quadratic function

$$L = 0.001189g^2 - 0.29243g + 17.949, \text{ with } g \in \{0...255\}, (\mathrm{R}^2 = 0.999). \tag{7.4}$$

7.2.5 Data Analysis

The statistical relationships among the experimental variables were analysed using an analysis-of-variance (ANOVA). An estimation of the number of image stimuli was made, in order to obtain the required degree of statistical power and sensitivity by the method discussed in Appendix C.

7.3 Results

This section discusses the results obtained from the experiments described in section 7.2.1. The data were explored by graphical means, and analysed by a within-subjects ANOVA to ascertain

statistical significance of the treatment effects. These effects are categorised as either *main effects* or higher order or *interaction effects*. The ANOVA tables are presented in section 7.3.4 on page 147. In some instances specific regression analyses were performed.

The main result of interest was the effect of the clutter radius on the subject's ability to detect targets. However, there were also other interesting and relevant effects that will be discussed and explored in differing degrees of detail.

7.3.1 Main Effects

The performance measure used for the analysis was the probability of detection, or hit-rate, which is defined, as elsewhere in this thesis, by equation (7.5). The probability of detection measure (hit-rate) was defined as:

$$p_{di} = \frac{1}{N} \sum_{j} W_{ij} \tag{7.5}$$

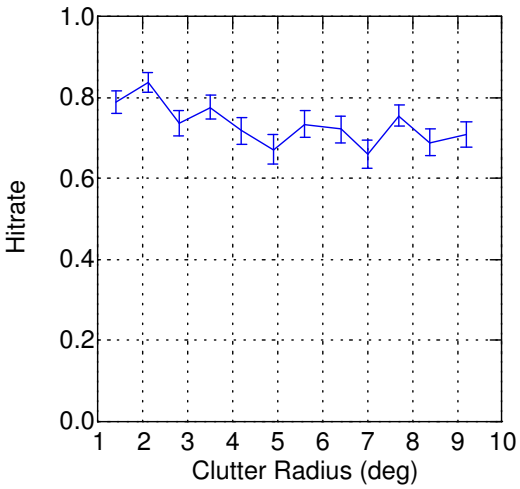
where W_{ij} is 0 for a miss and 1 for a hit for treatment i and subject j and N is the number of subjects. During the experimental sessions the observer's response time and confidence rating were also recorded as a potential performance measures. These procedures were described in section 7.2.3.

Response time proved not to be a very robust measure of performance in this experiment. With t_r as the dependent variable, the main effects for contrast and clutter parameter (δ) were highly significant, while the main effects for target radius and clutter radius were not significant (at the 0.05α level). From an analysis of individual subject response, it appeared that 60% of the subjects showed a causal relationship between the four factors and t_r, while the remaining subjects showed no effect. In contrast, all the main effects, for hit-rate (probability of detection, p_d) were highly significant; there existed a strong relationship between all factors and hit-rate in all subjects, which is summarised in figure 7.5. Hit-rate appears to be the most reliable and robust measure of performance in target detection (Ewing and Woodruff, 1996), see chapter 5. Response time appears to be valid only when the variation in the hit-rate is not large[9]. Therefore, since the hit-rate varied over a considerable range, only hit-rate will be discussed further.

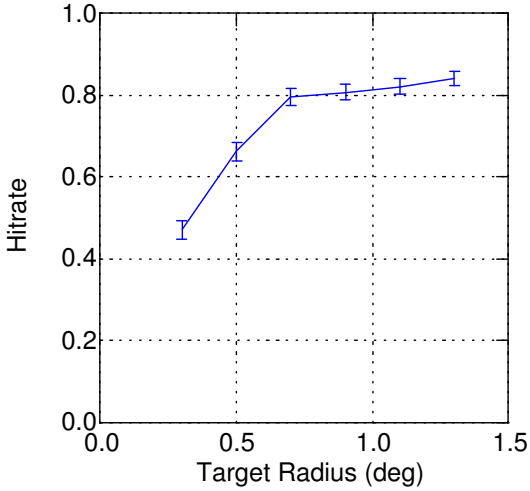
Figure 7.5 provides a graphical representation, including standard error bars, of the main effects for the independent variables. Figure 7.5(a) shows the main effect for hit-rate versus clutter radius. This curve is complex, but has a slight downward trend, with a possible levelling off after a clutter radius of about 5 degrees. Interaction effects may be coming into play here, adding to the complexity of the curve. To explore the trend of this effect, further ANOVA were performed, which included trend analysis; *i.e.* polynomial contrasts (Keppel, 1991a) were analysed. This will be discussed shortly.

Figure 7.5(b) shows the plot of hit-rate versus target radius. This graph indicates a strong effect of target radius on hit-rate for targets with a radius less than about 0.8 degrees of angle, where the effect for these small targets appears to be linear. As this plot indicates a dichotomy

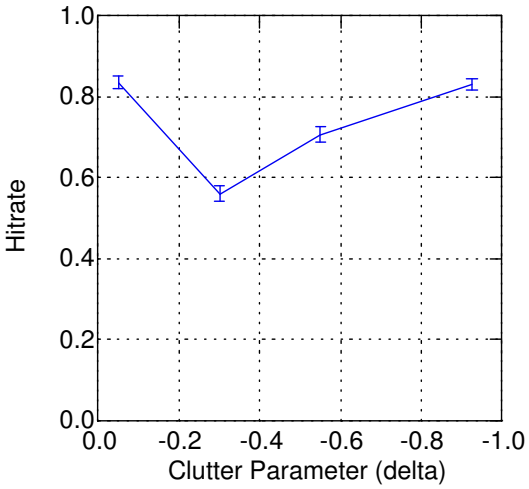
[9]This was discussed in section 5.3 in Chapter 5



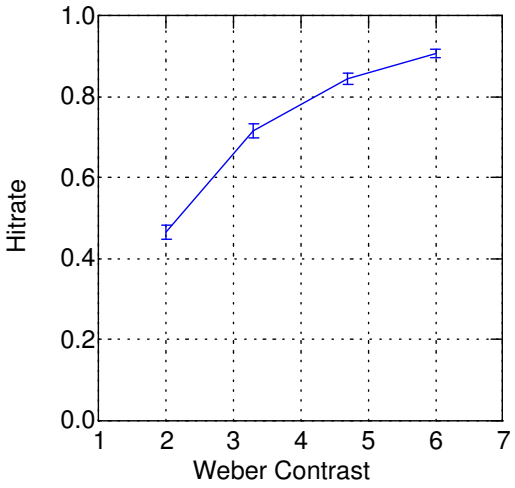
(a) Effect of Clutter Radius, $p < 0.001$.



(b) Effect of Target Radius, $p < 0.001$.



(c) Effect of Clutter Parameter, $p < 0.001$.

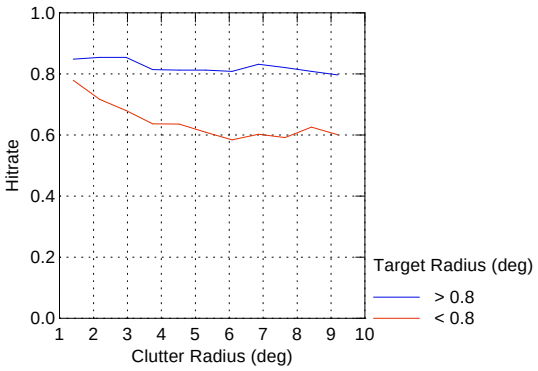


(d) Effect of Target Contrast, $p < 0.001$.

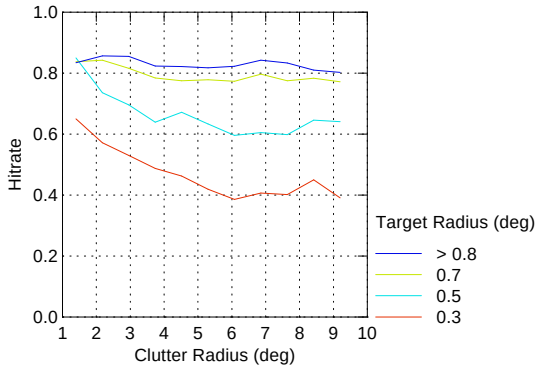
Figure 7.5: The effects that the independent variables have directly on the hit-rate.

of effects for small and large targets, further ANOVA were carried out separately for small and large targets (see tables 7.4 & 7.5). A graph of clutter radius versus hit-rate was produced, with separate plots for the targets that were classified as either small or large. This is shown in figure 7.6(a), and is discussed further in section 7.3.2.

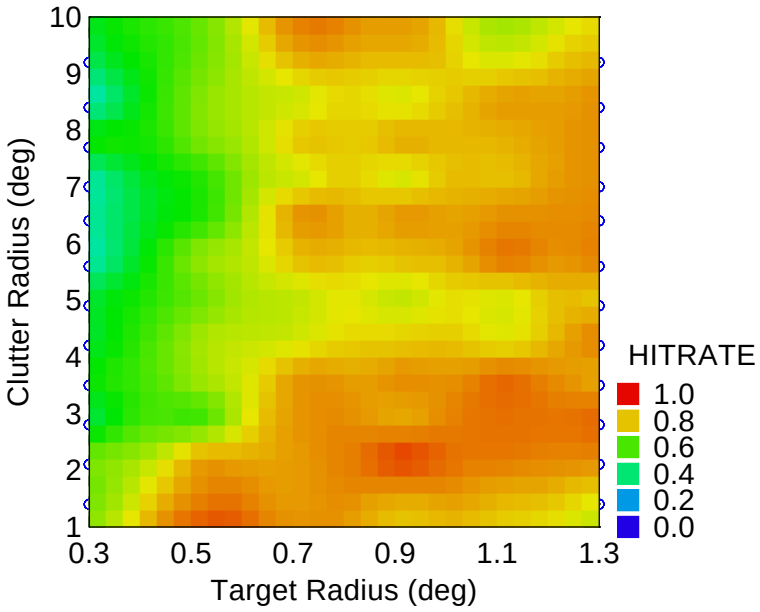
The effect of the clutter parameter on the hit-rate is presented in figure 7.5(c), with the four values of the clutter parameter (δ). The parameter δ has a range of -1 to 0 (see section 7.2.2), with δ near -1 indicating low correlation between pixels (*i.e.* the clutter approaches white noise in appearance), while a value for δ near 0, indicates a high correlation over a longer range between the pixels. A perusal of figure 7.5(c) shows that hit-rate drops for the intermediate values of δ, particularly at the value -0.3. This is discussed in section 7.3.2.



(a) Small and large targets.



(b) Individual small target sizes.



(c) Mosaic plot.

Figure 7.6: Interaction of clutter background size and target size.

Figure 7.5(d) shows the effect of target luminance contrast on hit-rate. As expected, there is an increase in hit-rate with increase in contrast, with the functional form of the relationship being almost linear. This indicates that, as intended, we are operating on or near the linear portion of the psychometric function.

7.3.2 Interactions

This being a 4 factor full factorial experiment, there were possible interactions up to 3$^{\text{rd}}$ order. This was done to allow the possibilty of exploring the higher order interactions, which was done. The ANOVA showed that higher order interactions were indeed statistically significant. These were explored graphically, and interactions above 1^{st} order were found to be too complex to interpret; *i.e.* the graphs were extremely convoluted, and it is hard to interpret data which requires up to 5 dimensions to represent. The 1^{st} order interactions may hide the details in higher order interactions, but represent them on average, and are much simpler to interpret. Therefore, only the 1^{st} order interactions are discussed here.

Target Size and Clutter Size

As indicated previously, further analyses of variance were performed for small and large targets, with the threshold between them being set at 0.8 degrees. These analyses are shown in tables 7.4 and 7.5 for small and large targets respectively, and are concerned with the interaction between target radius (small and large) and clutter radius. As mentioned earlier, these analyses included polynomial contrasts for trend analysis. Comparing the two analyses, we find that the main effect for target radius is significant in both cases, but much less so for large targets. The linear trend of clutter radius is also significant in both cases. However, a quadratic trend is evident for small targets, but not large targets. Cubic trends are not evident for either case, but it is noted that the contributions to the mean square (variance) for "Deviations" is significant. Therefore, although neither linear nor quadratic functions fully describe the plots, the trends are adequately described by these functions.

Figure 7.6 shows graphical representations of the interactions between clutter and target size. Because of considerable fluctuation in the data when referenced to clutter radius, as seen in the clutter radius main effect (figure 7.5(a)), the data plotted in figure 7.6 have been smoothed by an 11 point moving average filter[10]. The curves for the individual target sizes are included to show qualitative effects. However, due to there being insufficient data, these curves are not supported by statistical analysis, except that a significant interaction between clutter and target radius was demonstrated (table 7.3).

As can be seen in figure 7.6(a), there is no trend in the hit-rate for large targets, while there is a definite smooth trend downward with clutter radius for small targets (which is supported by the ANOVA just mentioned). This trend appears to level off at about 3.5° to 5.5° of clutter radius. Though figure 7.6(b) can be discussed in a qualitative way only, it indicates that the

[10] An N point moving average filter replaces the datum at its current output with the average of its N inputs. It then moves forward one place in the input data *etc.*

effect of clutter radius on hit-rate depends on target size. The gradient of the curve and the spatial extent of the clutter which affects hit-rate (and thus the amplitude) appears to depend on target size. This dependence would account for the significant interaction found between target radius and clutter radius.

Bearing in mind that these curves have been smoothed somewhat, figure 7.6(c) shows a pseudo 3-D plot of the actual raw data points. Here, the axes are clutter radius, target radius and hit-rate, which is coded (Z-modulated) by colour as shown in the legend. This is called a mosaic plot in Systat, which was used to produce it. On viewing this figure, there is an obvious structure, which appears, to use a geological metaphor, as a cliff-face above a green valley. The apparent contour formed from this "cliff-face" represents the interaction effects of clutter and target radius. According to this, there seems to be an interaction effect for clutter radii less than about 2.5° with target radii less than about 7°. For larger clutter and target radii, the effect on hit-rate seems fairly constant.

Target Size and Clutter Parameter

It was shown earlier in this section that the clutter parameter δ exhibited a main effect as shown graphically in figure 7.5(c). As was shown in Chapter 2, clutter must be referenced to a target. Therefore, the interaction between the clutter parameter and target radius should show some interesting properties. These interactions are shown graphically in figures 7.7(a) and 7.7(c). The curves in figure 7.7(a) exhibit the same shape as the δ main effect curve in figure 7.5(c), except for the smallest target radius of 0.3°. For this smallest target size tested, the range of δ, for effect on hit- rate, seems to have been extended. This same effect is evident in the 3-D plot in figure 7.7(c), with the extra dimension of target radius explicitly shown.

Contrast and Clutter Parameter

The interactions of contrast and the clutter parameter are represented in figures 7.7(b) & 7.7(d), with a standard 2-D plot in figure 7.7(b) and a 3-D plot in figure 7.7(d). The shapes of these curves are very similar to those for the target size * clutter parameter interaction, discussed in the last subsection. The implications of this are discussed in section 7.4.

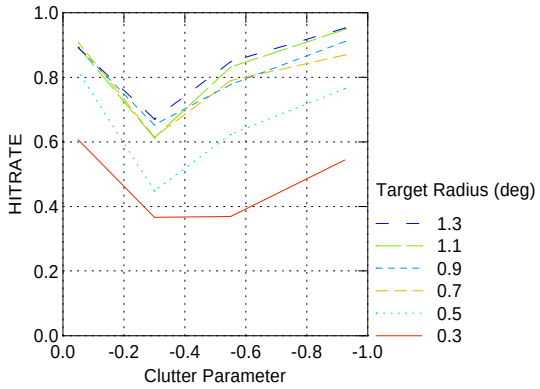
Contrast and Target Size

As discussed in Chapter 5 and section 5.2.2, the effect of contrast and target size on detection performance are inter-related; *viz.* the product of

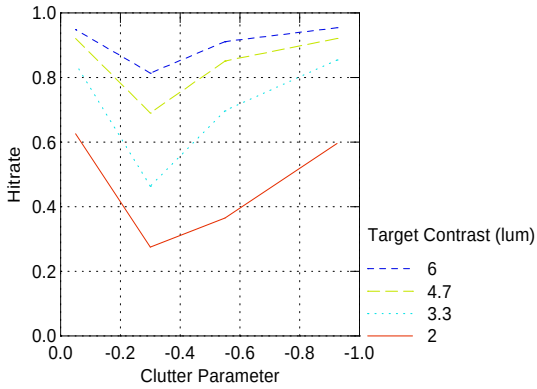
$$c \cdot r_t^2 = k, \text{ where } k \text{ is a constant,} \tag{7.6}$$

for small targets. This is a form of Ricco's Law (Barlow, 1958), which is often stated as

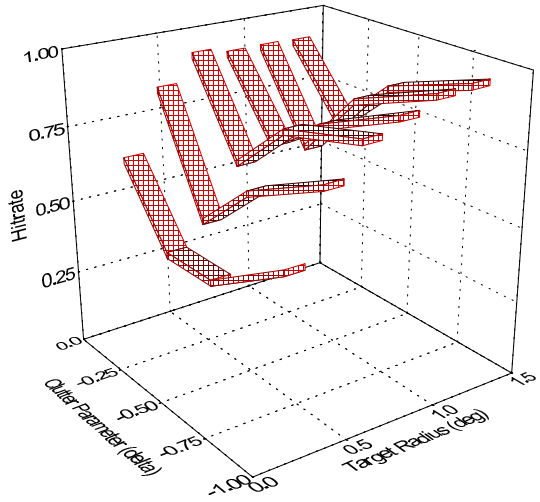
$$\frac{\Delta L}{L} \propto a^{-1}, \tag{7.7}$$



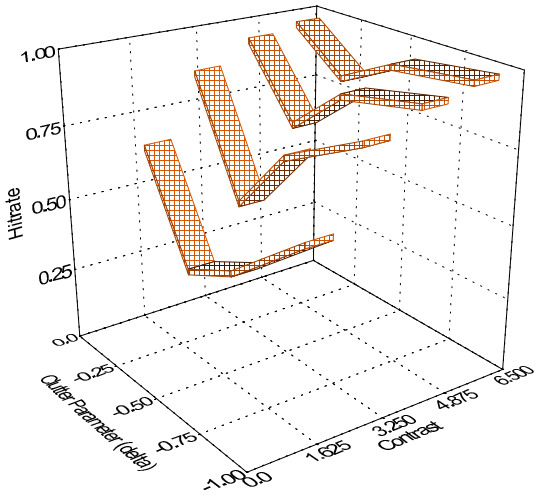
(a) $r_t * \delta$ interaction line plot.



(b) $c * \delta$ interaction line plot.



(c) $r_t * \delta$ interaction 3-D plot.



(d) $c * \delta$ interaction 3-D plot.

Figure 7.7: Clutter, contrast and clutter, target size interaction effects on hit-rate.

where ΔL is the incremental increase in luminance at threshold, of a disc of area a, over the background luminance L. Therefore a significant interaction is quite expected. Figures 7.8 and 7.9 illustrate these interaction effects. The graphs in figure 7.8 represent the effects of contrast on hit-rate for each target size. As expected, the hit-rate increases for increases in both contrast and target size, with target size "biasing" the hit-rate-versus-contrast curves to different levels (on the psychometric function). Although the curve for the smallest target radius of 0.3° looks perfectly linear, these curves become more non-linear with increases in target size. These effects are discussed in section 7.4. Figure 7.9 depicts the pseudo 3-D mosaic plot of hit-rate versus target radius and contrast. The plot is divided nicely into distinct constant hit-rate regions, bounded by hyperboloid curves. This illustrates that equation (7.6) is at least approximately true.

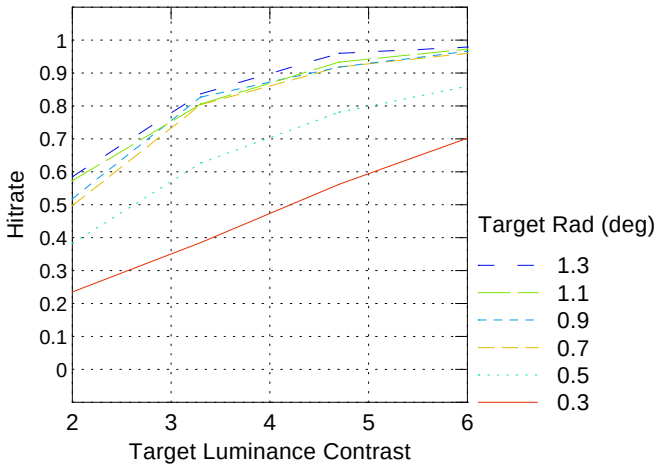


Figure 7.8: Target radius and contrast interaction effect on hit-rate expressed as line graphs. Each curve is a plot of hit-rate versus contrast for a given target angular radius.

7.3.3 Confidence Rating and Performance

As indicated in section 7.2, all subjects were required to record their confidence c_f in their decision as to whether they thought the stimulus in each trial contained a target. Confidence ratings were originally designed into the experimental software, in order to facilitate the production of receiver operating characteristic[11] (ROC) curves, but were not thought to add any appropriate information in this study. However, it is perhaps of interest to consider how the subject's confidence in their decision as to the existence of a target related to the objective reality, as indicated by the hit-rate p_d. This relationship is plotted in figure 7.10, with the data represented as points. The straight line is the line of best fit (linear regression), and provides an excellent fit ($R^2 = 0.955$). The regression equation is $p_d = 0.24c_f - 0.16$, or $c_f = 4.2p_d + 0.16$. This indicated that in this experiment, the subjective rating was a very good predictor of actual performance, but from the regression equation the subjects tended to under-estimate slightly their ability to detect targets.

[11]These are discussed in Chapters 3 and 9.

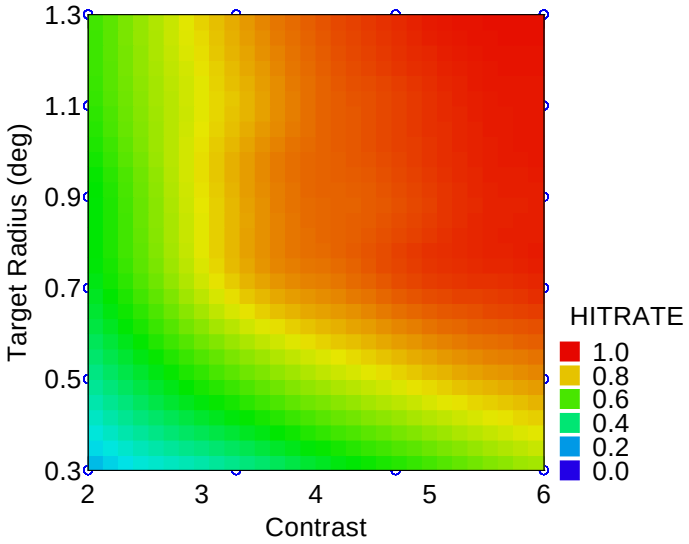


Figure 7.9: Target radius and contrast interaction effect on hit-rate expressed as a pseudo-3D plot. Hit-rate (value shown by colour) is plotted simultaneously against contrast and target angular radius.

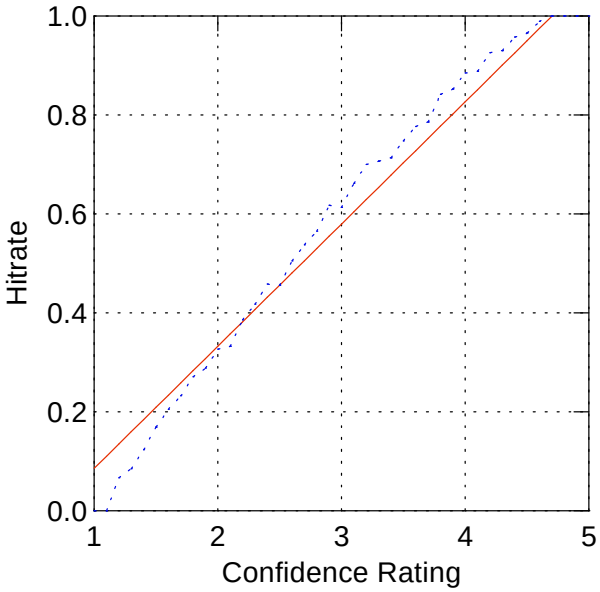


Figure 7.10: Regression of actual performance on to perceived performance ($R^2 = 0.955$).

7.3.4 ANOVA Tables

Four separate ANOVA tables are shown. Table 7.2 presents the results for the full analysis using response time (t_r) as the dependent variable. Table 7.3 presents the full analysis, using hit-rate as the dependent variable. Tables 7.4 and 7.5 show an analysis for targets $< 0.8°$ and targets $> 0.8°$ respectively.

Table 7.3 shows all the effects, including main effects and interactions, to be highly statistically significant ($p < 0.001$). This implies that the situation is very complex, and shows that the four main factors all contribute to visual task performance; *i.e.* none of the factors is redundant.

Table 7.2: Response time as the dependent variable.

Summary of all Effects for ANOVA - Response Time					
Effect	DF[12]	SS[13]	MS[14]	F Ratio	p-level
sub stratum	9	807.4878	89.7209	-	-
sub.contrast stratum					
contrast	3	62.6049	20.8683	8.82	<.001
Residual	27	63.9060	2.3669	-	-
sub.targ_rad stratum					
targ_rad	5	11.1332	2.2266	2.14	0.078
Residual	45	46.8690	1.0415	-	-
sub.clut_rad stratum					
clut_rad	11	5.9986	0.5453	1.84	0.056
Residual	99	29.2618	0.2956	-	-
sub.delta stratum					
delta	3	28.2729	9.4243	6.69	0.002
Residual	27	38.0205	1.4082	-	-
sub.contrast.targ_rad stratum					
contrast.targ_rad	15	16.7036	1.1136	3.67	<.001
Residual	135	41.0141	0.3038	-	-
sub.contrast.clut_rad stratum					
contrast.clut_rad	33	7.0693	0.2142	1.37	0.091
Residual	297	46.4128	0.1563	-	-
sub.targ_rad.clut_rad stratum					
targ_rad.clut_rad	55	19.7211	0.3586	1.77	<.001
Residual	495	100.2663	0.2026	-	-
sub.contrast.delta stratum					
contrast.delta	9	10.4418	1.1602	4.39	<.001
Residual	81	21.4000	0.2642	-	-
sub.targ_rad.delta stratum					
targ_rad.delta	15	9.4210	0.6281	3.90	<.001
Residual	135	21.7622	0.1612	-	-
sub.clut_rad.delta stratum					
clut_rad.delta	33	22.3555	0.6774	2.47	<.001
Residual	297	81.3782	0.2740	-	-
sub.contrast.targ_rad.clut_rad stratum					
contrast.targ_rad.clut_rad	165	30.0960	0.1824	1.04	0.361

continued on next page

continued from previous page					
Effect	DF	SS	MS	F Ratio	p-level
Residual	1485	260.8983	0.1757	1.07	-
sub.contrast.targ_rad.delta stratum					
contrast.targ_rad.delta	45	12.0620	0.2680	1.45	0.035
Residual	405	74.8021	0.1847	1.12	-
sub.contrast.clut_rad.delta stratum					
contrast.clut_rad.delta	99	20.5623	0.2077	1.22	0.077
Residual	891	151.1871	0.1697	1.03	-
sub.targ_rad.clut_rad.delta stratum					
targ_rad.clut_rad.delta	165	38.6534	0.2343	1.31	0.007
Residual	1485	265.0813	0.1785	1.09	-
sub.contrast.targ_rad.clut_rad.delta stratum					
contrast.targ_rad.clut_rad.delta	495	92.3190	0.1865	1.13	0.026
Residual	4455	732.1806	0.1644	-	-
Total	11519	3169.3427	-	-	-

Table 7.3: Hit-rate as the dependent variable.

Summary of all Effects for ANOVA - Hit-rate					
Effect	DF	SS	MS	F Ratio	p-level
sub stratum	9	2962.0118	329.1124	-	-
sub.contrast stratum					
contrast	3	5721.1028	1907.0343	235.65	<.001
Residual	27	218.4979	8.0925	-	-
sub.targ_rad stratum					
targ_rad	5	3193.6184	638.7237	98.74	<.001
Residual	45	291.0934	6.4687	-	-
sub.clut_rad stratum					
clut_rad	11	431.6444	39.2404	15.21	<.001
Residual	99	255.4007	2.5798	-	-
sub.delta stratum					
delta	3	2627.0937	875.6979	102.26	<.001
Residual	27	231.2222	8.5638	-	-
sub.contrast.targ_rad stratum					
contrast.targ_rad	5	95.5774	6.3718	3.58	<.001
Residual	135	239.9885	1.7777	-	-
sub.contrast.clut_rad stratum					
contrast.clut_rad	33	104.5181	3.1672	4.16	<.001
Residual	297	226.0062	0.7610	-	-
sub.targ_rad.clut_rad stratum					
targ_rad.clut_rad	55	581.2316	10.5678	10.56	<.001
Residual	495	495.4316	1.0009	-	-
sub.contrast.delta stratum					
contrast.delta	9	254.6382	28.2931	10.36	<.001
Residual	81	221.1431	2.7302	-	-
sub.targ_rad.delta stratum					
targ_rad.delta	15	226.4281	15.0952	11.98	<.001

continued on next page

continued from previous page					
Effect	DF	SS	MS	F Ratio	p-level
Residual	135	170.0476	1.2596	-	-
sub.clut_rad.delta stratum					
clut_rad.delta	33	1144.7437	34.6892	21.65	<.001
Residual	297	475.9819	1.6026	-	-
sub.contrast.targ_rad.clut_rad stratum					
contrast.targ_rad.clut_rad	165	267.3392	1.6202	1.90	<.001
Residual	1485	1266.4698	0.8528	1.09	-
sub.contrast.targ_rad.delta stratum					
contrast.targ_rad.delta	45	343.5399	7.6342	7.06	<.001
Residual	405	437.9288	1.0813	1.38	-
sub.contrast.clut_rad.delta stratum					
contrast.clut_rad.delta	99	321.3576	3.2460	3.45	<.001
Residual	891	837.5694	0.9400	1.20	-
sub.targ_rad.clut_rad.delta stratum					
targ_rad.clut_rad.delta	165	1142.0719	6.9216	8.34	<.001
Residual	1485	1232.4108	0.8299	1.06	-
sub.contrast.targ_rad.clut_rad.delta stratum					
contrast.targ_rad.clut_rad.delta	495	721.2267	1.4570	1.86	<.001
Residual	4455	3487.5962	0.7828	-	-
Total	11519	30224.9319	-	-	-

Table 7.4: ANOVA Table for small targets.

Summary of Effects for ANOVA (target < 0.8°)					
Effect	DF	SS	MS	F Ratio	p-level
sub stratum	9	136.8342	15.2038	-	-
sub.contrast stratum					
contrast	3	181.8339	60.6113	122.41	<.001
Residual	27	13.3693	0.4952	-	-
sub.targ_rad stratum					
targ_rad	2	101.8792	50.9396	60.24	<.001
Residual	18	15.2215	0.8456	-	-
sub.clut_rad stratum					
clut_rad	11	31.3130	2.8466	16.22	<.001
Lin	1	13.2304	13.2304	75.38	<.001
Quad	1	7.6891	7.6891	43.81	<.001
Cub	1	0.0408	0.0408	0.23	0.631
Quart	1	0.0192	0.0192	0.11	0.742
Deviations	7	10.3335	1.4762	8.41	<.001
Residual	99	17.3762	0.1755	-	-
sub.delta stratum					
delta	3	79.2019	26.4006	42.24	<.001
Residual	27	16.8762	0.6250	-	-
sub.contrast.targ_rad stratum					
contrast.targ_rad	6	4.4417	0.7403	2.87	0.017
continued on next page					

continued from previous page

Effect	DF	SS	MS	F Ratio	p-level
Residual	54	13.9437	0.2582	-	-
sub.contrast.clut_rad stratum					
contrast.clut_rad	33	4.2974	0.1302	1.28	0.147
contrast.Lin	3	0.7153	0.2384	2.34	0.073
contrast.Quad	3	0.9727	0.3242	3.19	0.024
contrast.Cub	3	0.0584	0.0195	0.19	0.902
Deviations	24	2.5509	0.1063	1.04	0.409
Residual	297	30.2286	0.1018	-	-
sub.targ_rad.clut_rad stratum					
targ_rad.clut_rad	22	16.8625	0.7665	6.73	<.001
targ_rad.Lin	2	2.4306	1.2153	10.68	<.001
targ_rad.Quad	2	1.1300	0.5650	4.96	0.008
targ_rad.Cub	2	0.3024	0.1512	1.33	0.267
Deviations	16	12.9994	0.8125	7.14	<.001
Residual	198	22.5368	0.1138	-	-
sub.contrast.delta stratum					
contrast.delta	9	10.8641	1.2071	8.94	<.001
Residual	81	10.9398	0.1351	-	-
sub.targ_rad.delta stratum					
targ_rad.delta	6	6.0403	1.0067	7.71	<.001
Residual	54	7.0535	0.1306	-	-
sub.clut_rad.delta stratum					
clut_rad.delta	33	55.3127	1.6761	13.16	<.001
Lin.delta	3	17.5995	5.8665	46.05	<.001
Quad.delta	3	6.8769	2.2923	17.99	<.001
Cub.delta	3	1.6700	0.5567	4.37	0.005
Deviations	24	29.1662	1.2153	9.54	<.001
Residual	297	37.8384	0.1274	-	
sub.contrast.targ_rad.clut_rad stratum					
contrast.targ_rad.clut_rad	66	7.6333	0.1157	1.17	0.181
contrast.targ_rad.Lin	6	0.6799	0.1133	1.14	0.335
contrast.targ_rad.Quad	6	0.7492	0.1249	1.26	0.273
Deviations	54	6.2042	0.1149	1.16	0.209
Residual	594	58.8146	0.0990	0.96	-
sub.contrast.targ_rad.delta stratum					
contrast.targ_rad.delta	18	5.4333	0.3019	3.30	<.001
Residual	162	14.8201	0.0915	0.89	-
sub.contrast.clut_rad.delta stratum					
contrast.clut_rad.delta	99	19.3130	0.1951	1.97	<.001
contrast.Lin.delta	9	3.6266	0.4030	4.07	<.001
contrast.Quad.delta	9	0.9123	0.1014	1.02	0.419
Deviations	81	14.7739	0.1824	1.84	<.001
Residual	891	88.2373	0.0990	0.96	-
sub.targ_rad.clut_rad.delta stratum					
targ_rad.clut_rad.delta	66	49.5514	0.7508	7.47	<.001
targ_rad.Lin.delta	6	7.8419	1.3070	13.01	<.001

continued on next page

continued from previous page					
Effect	DF	SS	MS	F Ratio	p-level
targ_rad.Quad.delta	6	11.0524	1.8421	18.33	<.001
Deviations	54	30.6571	0.5677	5.65	<.001
Residual	594	59.6882	0.1005	0.97	-
sub.contrast.targ_rad.clut_rad.delta stratum					
	1980	204.0799	0.1031	-	-
Total	5759	1321.8359	-	-	-

Table 7.5: ANOVA Table for large targets.

Summary of Effects for ANOVA (target > 0.8°)					
Effect	DF	SS	MS	F Ratio	p-level
sub stratum	9	32.70069	3.63341	-	-
sub.contrast stratum					
contrast	3	151.58403	50.52801	54.83	<.001
Residual	27	24.88125	0.92153	-	-
sub.targ_rad stratum					
targ_rad	2	1.04410	0.52205	5.00	0.019
Residual	18	1.87951	0.10442	-	-
sub.clut_rad stratum					
clut_rad	11	10.50347	0.95486	8.46	<.001
Lin	1	1.30679	1.30679	11.58	<.001
Quad	1	0.00223	0.00223	0.02	0.889
Cub	1	0.00161	0.00161	0.01	0.905
Deviations	8	9.19284	1.14910	10.18	<.001
Residual	99	11.17014	0.11283	-	-
sub.delta stratum					
delta	3	71.15764	23.71921	47.97	<.001
Residual	27	13.34931	0.49442	-	-
sub.contrast.targ_rad stratum					
contrast.targ_rad	6	0.87118	0.14520	2.62	0.027
Residual	54	2.99688	0.05550	-	-
sub.contrast.clut_rad stratum					
contrast.clut_rad	33	13.09514	0.39682	5.95	<.001
contrast.Lin	3	1.87534	0.62511	9.37	<.001
contrast.Quad	3	1.50353	0.50118	7.51	<.001
contrast.Cub	3	0.52202	0.17401	2.61	0.052
Deviations	24	9.19425	0.38309	5.74	<.001
Residual	297	19.81458	0.06672	-	-
sub.targ_rad.clut_rad stratum					
targ_rad.clut_rad	22	7.74757	0.35216	4.18	<.001
targ_rad.Lin	2	0.98327	0.49163	5.84	0.003
targ_rad.Quad	2	1.00945	0.50473	6.00	0.003
targ_rad.Cub	2	0.15406	0.07703	0.92	0.402
Deviations	16	5.60079	0.35005	4.16	<.001
Residual	198	16.66215	0.08415	-	-

continued on next page

continued from previous page

Effect	DF	SS	MS	F Ratio	p-level
sub.contrast.delta stratum					
contrast.delta	9	33.62986	3.73665 1	8.22	<.001
Residual	81	16.61319	0.20510	-	-
sub.targ_rad.delta stratum					
targ_rad.delta	6	1.63090	0.27182	4.59	<.001
Residual	54	3.19549	0.05918	-	-
sub.clut_rad.delta stratum					
clut_rad.delta	33	55.3127	1.6761	13.16	<.001
Lin.delta	3	17.5995	5.8665	46.05	<.001
Quad.delta	3	6.8769	2.2923	17.99	<.001
Cub.delta	3	1.6700	0.5567	4.37	0.005
Deviations	24	29.1662	1.2153	9.54	<.001
Residual	297	37.8384	0.1274	-	
sub.contrast.targ_rad.clut_rad stratum					
contrast.targ_rad.clut_rad	66	6.98715	0.10587	1.57	0.004
contrast.targ_rad.Lin	6	0.42869	0.07145	1.06	0.387
contrast.targ_rad.Quad	6	0.96879	0.16147	.39	0.027
Deviations	54	5.58967	0.10351	1.53	0.011
Residual	594	40.14479	0.06758	1.01	-
sub.contrast.targ_rad.delta stratum					
contrast.targ_rad.delta	18	4.89826	0.27213	3.58	<.001
Residual	162	12.31701	0.07603	1.13	-
sub.contrast.clut_rad.delta stratum					
contrast.clut_rad.delta	99	22.79097	0.23021	3.18	<.001
contrast.Lin.delta	9	9.85023	1.09447	15.14	<.001
contrast.Quad.delta	9	3.18443	0.35383	4.89	<.001
Deviations	81	9.75631	0.12045	1.67	<.001
Residual	891	64.42431	0.07231	1.08	-
sub.targ_rad.clut_rad.delta stratum					
targ_rad.clut_rad.delta	66	15.46076	0.23425	3.49	<.001
targ_rad.Lin.delta	6	0.69141	0.11524	1.72	0.115
targ_rad.Quad.delta	6	0.56937	0.09490	1.41	0.207
Deviations	54	14.19998	0.26296	3.92	<.001
Residual	594	39.87951	0.06714	1.00	-
sub.contrast.targ_rad.clut_rad.delta stratum					
	1980	132.95139	0.06715	-	-
Total	5759	838.08264	-	-	-

7.4 Discussion

The graphs as shown in figure 7.7(a) are most interesting. Why does the relative hit-rate drop for all target sizes at $\delta = -0.3$, and for a target radius of 0.3° at $\delta = -0.55$? It might be thought that the correlation length, imposed on the stimuli by a value for $\delta = -0.3$, was probably most similar to the range of target sizes used, thereby producing a confounding effect due to clutter granularity being about the same size as the targets. But consider figure 7.4, which shows the 2-D profiles of the circularly symmetric correlation function (equation (7.3)) for the displayed images. If we compare the range of target radii of 4 - 18 pixels, with the full-width-half-maximum (FWHM) value for the correlation length, we would expect that the hit-rate would decrease for increasing (*i.e.* more positive) δ, as the mean target radius approached the correlation length. This is not observed, as shown in figure 7.7(a), and therefore does not offer an explanation for the experimental curves.

It seems more likely that, since the $c * \delta$ interaction curves (figure 7.7(b)) produced a very similar set of curves to those for the $t_r * \delta$ interaction, this phenomenon must be related to a generalised *virtual contrast* or signal-to-noise ratio. That is, there are visual parameters that are not directly contrast related, but impact on perception as contrast effects. This approach is explored in the remainder of this section.

Consider an image[15], $L(x, y)$ which is a GRF, as defined in section 7.2.2, with co-variance function

$$C(\alpha, \beta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [L(x + \alpha) - \mu(x)][L(y + \beta) - \mu(y)] \, dxdy, \tag{7.8}$$

where μ is a mean, and $L(x, y)$ is assumed to be stationary and ergodic (Yaglom, 1987). This can be considered, without loss of generality, in one dimension, especially in our case with isotropic image statistics. If we also assume zero mean, then (7.8) becomes

$$C(\alpha) = \int_{-\infty}^{\infty} L(x + \alpha)L(\alpha) \, dx. \tag{7.9}$$

This is in fact the auto-covariance function, since this describes the correlation between points within a single image. By substituting $\alpha = -x$, (7.9) becomes

$$f(\alpha) = \int_{-\infty}^{\infty} L(\alpha - x)L(x) \, dx, \tag{7.10}$$

which is the convolution integral (Gonzalez and Wintz, 1987b). In general, convolution is defined as

$$f(x) * g(x) = \int_{-\infty}^{\infty} f(\alpha - x)g(x) \, dx, \tag{7.11}$$

where $*$ is the convolution operator. This is often used in the engineering analysis of linear shift-invariant systems (Karbowiak, 1969a), as (7.11) represents a filtering operation. It will also be used in a model of the HVS to be discussed shortly.

[15]The image is considered in the continuous domain, even though it is at this stage digital.

To facilitate further discussion, a simplified diagram, representing a one dimensional mapping of the displayed image on the viewer's retina, is shown in figure 7.11, where x is a unit-less distance variable.

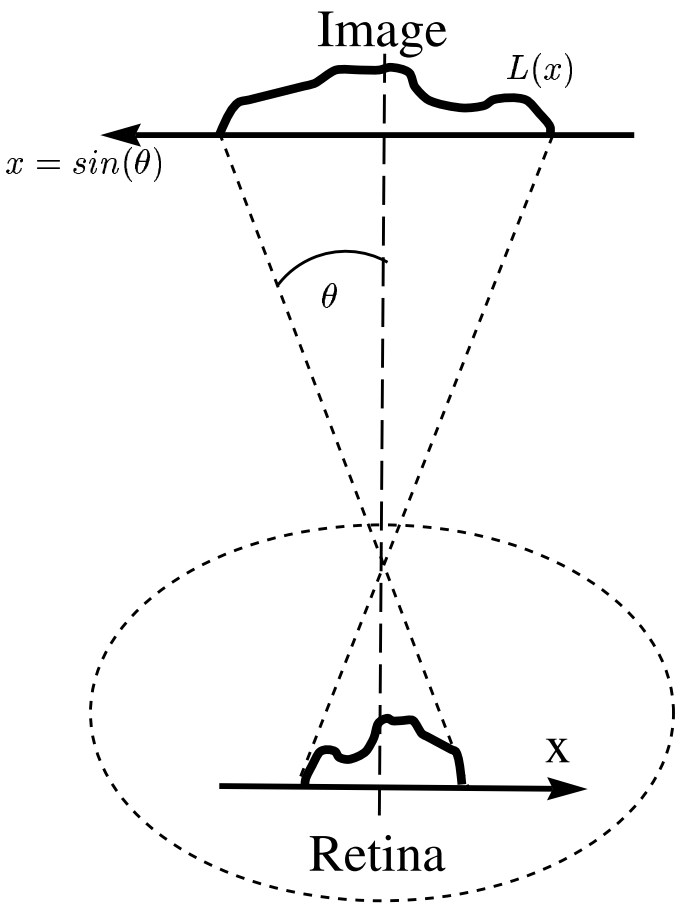


Figure 7.11: Mapping of viewed image to retina in 1-D.

For vision out to about 30° of periphery, the HVS has been modelled as a linear, shift-invariant system (Laming, 1986; Thibos, 1989), which is a useful model for development here. Thibos (1989) showed that the output from a foveal or near-foveal receptive field (Overington, 1982) is given by

$$r(u) = w(x) * L_r(x) = \int_{rf} w(u-x)L_r(x)\,dx, \tag{7.12}$$

where $w(x)$ is a weighting function over the spatial response of the receptive field and $L_r(x) = p(x) * L(x)$ is the luminance function at the retina, where $p(x)$ is the point-spread-function[16] of the eye.

Now consider the present experiment, with a displayed image consisting of a background clutter image $L(x)$, characterised in general terms, with the co-variance function defined in (7.8)

[16]See Chapter 2 section 2.2.4.

and which is determined by the value of δ, the clutter parameter. The target $L_t(x)$ was a disc produced by a adding a luminance increment ΔL_t to the background; *i.e.* the input image to the eye was

$$L''(x) = L'(x) * f_\delta(x) + L_t(x), \tag{7.13}$$

where $L(x) = L'(x) * f_\delta(x)$ and $f_\delta(x)$ is the filtering function defined by (7.10) and is applied to a conceptual, uncorrelated image $L'(x)$, to produce the actual observed background image $L(x)$. Combine (7.12) and (7.13) to obtain

$$r(u) = w(x) * p(x) * (L'(x) * f_\delta(x) + L_t(x)), \tag{7.14}$$

$$= w(x) * p(x) * L'(x) * f_\delta(x) + w(x) * p(x) * L_t(x), \tag{7.15}$$

since convolution is a linear operation. The effects of $f_\delta(x)$ will "swamp" the effects of $w(x)*p(x)$. Anyway, $w(x) * p(x)$ is relatively fixed, since viewing geometry and conditions are constant. Therefore,

$$r(u) \approx L(x) * f_\delta(x)' + L_t(x), \tag{7.16}$$

where $f_\delta(x)'$ is the equivalent function which, when convolved with a given $L(x)$, yields the same effect as $L'(x) * f_\delta(x)$.

Now, this is related to the effects represented in figure 7.7(a). In the case where the clutter parameter δ approaches -1, the clutter image becomes less correlated; *i.e.* approaches white noise. Then the auto-correlation function of the image approaches a delta-function ($\delta(x)$[17]) (Karbowiak, 1969b), where the convolution of a δ-function with another function produces this same function; *i.e.*

$$f(\alpha) = \int_{-\infty}^{\infty} g(x)\delta(\alpha - x)\, dx = g(\alpha). \tag{7.17}$$

Therefore, for the clutter parameter $\delta = -0.925$, (7.16) becomes

$$r(u) \approx L(x) + L_t(x), \tag{7.18}$$

thereby maintaining the input image signal-to-noise ratio or contrast; *i.e.* the virtual contrast is approximately the physical contrast. This situation is diagrammatically represented in figure 7.12 (a). At the other end of the scale, for $\delta \to 0$, the function $f_\delta(x)$ spreads out the energy in $L(x)$, also producing a relatively high virtual contrast (figure 7.12 (c)). However, for intermediate values of δ, $f_\delta(x) * L(x)$ is less spread out and of higher amplitude, thereby reducing the virtual contrast between target and background (figure 7.12 (b)).

The region for full spatial integration of the retinal luminance function $L_r(x)$ extends out (off axis) to about 0.5° of visual angle according to Marlow and Laming (Marlow, 1958; Laming, 1986); *i.e.* the region for which Ricco's Law (7.7) is obeyed. Apparently, outside this region, a

[17] This δ is not to be confused with the clutter parameter.

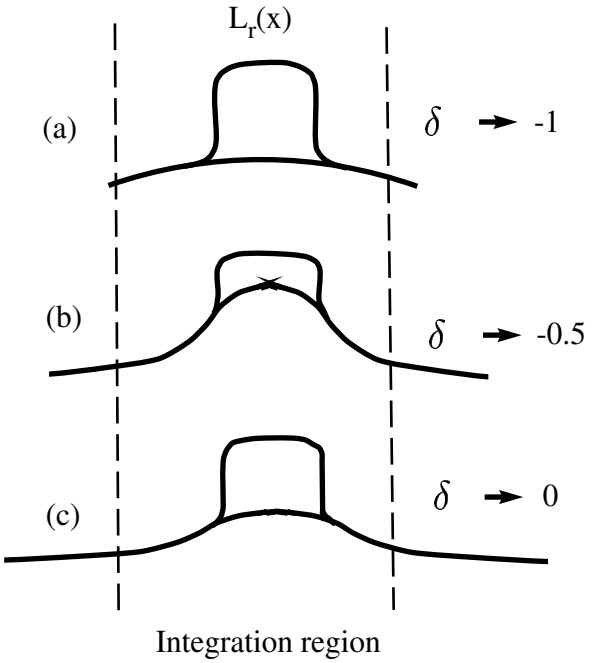


Figure 7.12: Virtual contrast at retina (1-D). Here $L_r(x)$ is the luminance function of the stimulus as 'seen' by the retina, and x is a dimensionless measure of distance perpendicular to the incoming light (see figure 7.11).

form of square law summation occurs (Laming, 1986), where, for stimuli persisting longer than 0.93 seconds, Ricco's Law is modified to

$$\frac{\Delta L}{L} \propto a^{-}\frac{1}{4}. \tag{7.19}$$

This would explain the flattening of the curve (in figure 7.7(a)) for 0.3° radius targets, since up till 0.5°, the background luminance would contribute equally to $r(u)$ for equivalent levels of the target luminance. However, targets larger than 0.5° radius have the background luminance contribution falling off according to equation (7.19). Therefore, the 0.3° target has a relatively lower virtual contrast and an associated loss in sensitivity to δ value.

The model just discussed can be viewed from a spatial frequency point of view, since it is known that the convolution of two functions, say $f(x) * g(x)$, in the x domain is equivalent to $F(\nu) \cdot G(\nu)$ in the frequency (ν) domain, where $F(\nu)$ and $G(\nu)$ are the Fourier transforms of $f(x)$ and $g(x)$ respectively[18] (Yaglom, 1987; Karbowiak, 1969a). Also the Fourier transform of the auto-covariance function $C(x)$ is the spectral density $S(x)$, which is the equivalent spectral characterisation of the function. These apply equally in the general 2-D case for images (Gonzalez and Wintz, 1987b). This is an appropriate way to analyse vision since it is well known that the HVS incorporates spatial frequency channels (Wilson, 1995).

Now the phenomena discussed are re-presented in the spectral context, although only briefly and qualitatively. If we consider figure 7.7(a), as δ goes from -1 to 0, the image is effectively low pass filtered, with the frequency content becoming lower in frequency. Consider now the

[18]See Appendix I for a mathematical development of digital filtering theory.

clutter image with $\delta = -0.925$, where the image is dominated by high frequencies, due to sharp transitions between the relatively uncorrelated pixels. The insertion of a target introduces lower frequency components, cueing the HVS to the target area, though local high frequency (edge) effects probably localise the target (Burr et al., 1979), so that here the virtual contrast is high. At the other end of the range, where δ is near 0, the background image is highly correlated, and therefore dominated by low frequencies. The insertion of a target, which is small compared to the correlation length, introduces relatively high spatial frequencies, again producing high virtual contrast. However, at intermediate values of δ, the frequency content of the background image must overlap the frequency range introduced by the targets, causing lower virtual contrast and resulting in lower hit-rates.

Under this hypothesis, when applied to the main effect of clutter radius on hit-rate, we would expect to see little or no fall-off for δ near -1 and the largest fall-off in hit-rate with δ near 0. To test this, another plot was produced of hit-rate versus clutter radius, but with separate plots for each δ value. This is shown in figure 7.13, where the data has been moving-average-filtered as before. This graph shows hit-rate to be independent of clutter radius for $\delta = -0.925$ and the greatest fall-off with $\delta = -0.05$, as expected. At the intermediate values of δ, the situation is slightly more complex, but an intermediate fall-off in hit-rate with clutter radius is evident.

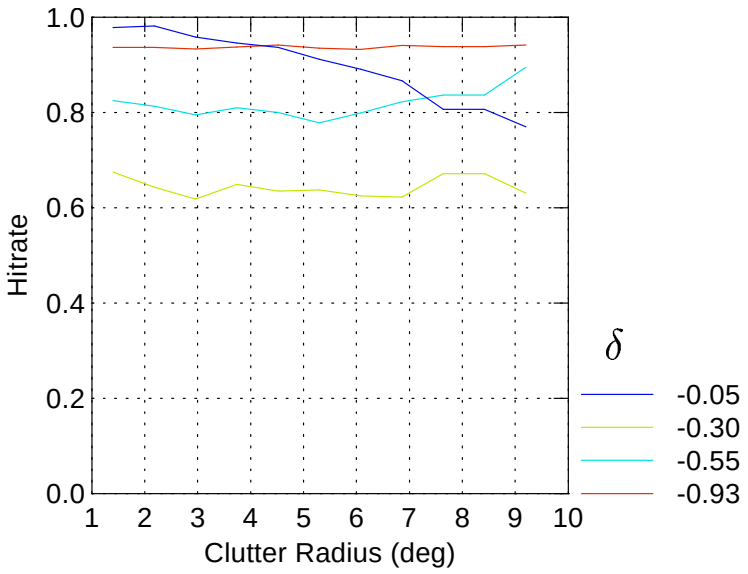


Figure 7.13: Clutter radius interaction with δ value.

A plausible explanation can now be given as to why the hit-rate for larger targets is less affected by clutter radius than that for small targets. It seems that the size of the region of stimulus integration is determined by the correlation function of the clutter. However, given this, the virtual contrast is then determined by the target size. For targets smaller than the area for full integration (about 0.5°), part of the clutter stimulus also falls within this region. Therefore, the rate of fall-off in hit-rate for small targets, with increasing clutter radius, is greater compared to larger targets, which force the clutter background luminance into a region, that is integrated

at a lower rate. This, in turn, reduces the sensitivity of subjective hit-rate to clutter radius with larger targets.

7.5 Conclusions

It was found that the size of the local clutter region around a target has a strong effect on the probability of detection of that target and that this is affected by regions much larger than twice the target size, as routinely used in the literature for setting clutter metric regions of support. It was also discovered that this effect was much stronger for targets subtending less than 0.8 degrees of visual angle than for larger targets. In the case of the former, the fall-off in human visual performance with clutter region size was approximately quadratic compared to a slight linear fall-off for larger targets.

A simple model was presented explaining these phenomena, indicating that the auto-covariance function characterising the clutter is the main determinant of the size of the region of local clutter, but is reduced for larger targets. The large regions for stimulus integration assumed in this model are much larger than the areas for single receptive fields, but have been shown to exist by other researchers, as discussed in section 7.4. The work reported here did not elucidate the detailed stimulus interactions across multiple receptive fields and further work needs to be done in order to fully understand the mechanisms.

This work considered only a narrow class of clutter and simple target type. Even for this situation, more work needs to be done in order to fully understand the mechanisms involved.

Part 3

THE APPLICATION OF IMAGE MEASURES TO THE PREDICTION OF HUMAN OBSERVER PERFORMANCE

Summary: This final part of the thesis describes two studies investigating the application of image quality and clutter metrics to real[19] world imagery. An image metric, for measuring the properties of images, to optimise the processing applied to them, was developed and tested. This metric was found to be quite effective and it appeared to agree with subjective judgement, though this was not comprehensively tested. A further, very large study, investigated the effects of clutter on human target detection performance. A well cited clutter metric was put to its first rigorous test on real imagery and using knowledge gained from Part 2 of this thesis. With the class of imagery used, this proved to be a good metric for predicting human visual performance in clutter, and it validated other findings of this thesis. Finally, the conclusions reached from the work described in this thesis are summarised, and suggestions are made for further work.

[19] Almost every study described in this thesis used real images, but here the application had real practical outcomes.

Chapter 8

The Gradient Energy Measure

Summary: *This chapter is a summary of part of my work carried out at the Department of Nuclear Medicine at the Queen Elizabeth Hospital, Adelaide. This work was motivated by my interest in image quality measures, which led me to consider the effects of image processing parameters on the clinical value of Single Photon Emission Computed Tomography (SPECT) images. The overall aim was to develop an automatic system for optimal image filter parameter adjustment.*

A measure, called the gradient energy measure (GEM), for quantifying the effect of filtering on SPECT images was developed and evaluated. This proved to be a reliable measure of image smoothing, and noise level, which in preliminary studies agreed with human perception.

8.1 Introduction

Image processing techniques are used in medical imaging to enhance, restore and code images in such a way that the quality of the processed images is maintained or improved. However, when these techniques are employed, the question arises as to how they may be evaluated. This in turn, implies that a method is needed to quantify the quality of the processed image. In this context, the mean square error (MSE) is very commonly used as a measure of quality. However, the MSE measures only one aspect of image quality and, as has been shown, both in the literature[1] and experimentally in Chapter 4, it does not correlate well with human perception. Theoretical issues concerning such measures were discussed, both in Chapter 1 and Chapter 2. In contrast, here we are concerned with the practical application of such a measure to a real problem.

The work described in this chapter was the initial phase of a program of research to improve the diagnostic quality or utility of brain images, by optimising image processing parameters with respect to expert human observers (Nuclear Medicine Physicians). This was done by developing an image measure, which was sensitive to appropriate filtering parameters, and (though not the main aim) which correlated with human subjective[2] evaluation of these filtering operations. The

[1] See Chapter 2 for a discussion on this literature.

[2] Interval scale ratings obtained from 4 expert observers.

remainder of this chapter describes the development and evaluation of this measure. However, we begin with some background discussion on its application.

8.1.1 Emission Computed Tomography

Several medical imaging modalities use a technique known as tomography. This is the cross-sectional imaging of a patient from either transmission, emission or reflection data, collected from many viewing angles around the patient; *i.e.* a 3-D image of the patient's body is built up from many (1-D) views.

In conventional x-ray tomography, physicians use the attenuation coefficient of tissue to infer diagnostic information about the patient. Emission computed tomography (ECT), on the other hand, uses the decay of radioactive isotopes to image the distribution of the isotope as a function of time. These isotopes may be administered to the patient, in the form of radio-pharmaceuticals, either by injection or by inhalation. Thus, for example, by administering a radioactive isotope by inhalation, ECT can be used to trace the path of the isotope through the lungs and the rest of the body.

Radioactive isotopes are characterised by the emission of gamma-ray photons or positrons, both products of nuclear decay. The concentration of such an isotope in any cross-section changes with time due to radioactive decay, flow, and biochemical kinetics within the body. This implies that all the data for one cross-sectional image must be collected in a time interval that is short compared to the time constant associated with the changing concentration. However, this aspect also provides ECT with its greatest potential and utility in diagnostic medicine: by analysing the images taken at different times for the same cross section, we can determine the functional state of various organs in a patient's body.

ECT is of two types: Single Photon Emission Computed Tomography (SPECT) and Positron Emission Tomography (PET). The word single in the former refers to the product of the radioactive decay, a single photon, while in PET, the decay produces a single positron. After travelling a short distance, the positron comes to rest and combines with an electron. The annihilation of the emitted positron results in two gamma-ray photons travelling in opposite directions. Only SPECT is discussed here.

8.1.2 Single Photon Emission Tomography

Figure 8.1 shows a cross-section of a patient with a distributed source emitting gamma-ray photons. For the purpose of imaging, any element of this source that is very small, compared with the whole field of view, may be considered to be an isotropic source of gamma-rays. The number of gamma-ray photons emitted per second by such an element is proportional to the concentration of the source at that point. If we assume that the collimator in front of the detector has infinite collimation[3], it will accept only those photons that travel toward it in the parallel

[3]Infinite collimation, in practice, would imply an infinitely long time to make a statistically meaningful observation.

ray-bundle R_1R_2. The total number of photons recorded by the detector in a "statistically meaningful" time interval is then proportional to the total concentration of the emitter along the line defined by R_1R_2. This summation of photon counts along this line is known as a "ray integral"[4]. By moving the detector-collimator assembly to an adjacent position laterally, one may determine this integral for another ray parallel to R_1R_2. After one such scan is completed, generating one projection, one may either rotate the patient or the detector-collimator assembly and generate other projections. Under ideal conditions, it should be possible to generate the projection data required for the usual reconstruction algorithms. It is beyond the scope of this thesis to discuss these algorithms here, but the book by Kak and Slaney (Kak and Slaney, 1988) gives a readable, though necessarily mathematical, description of the commonly used reconstruction techniques.

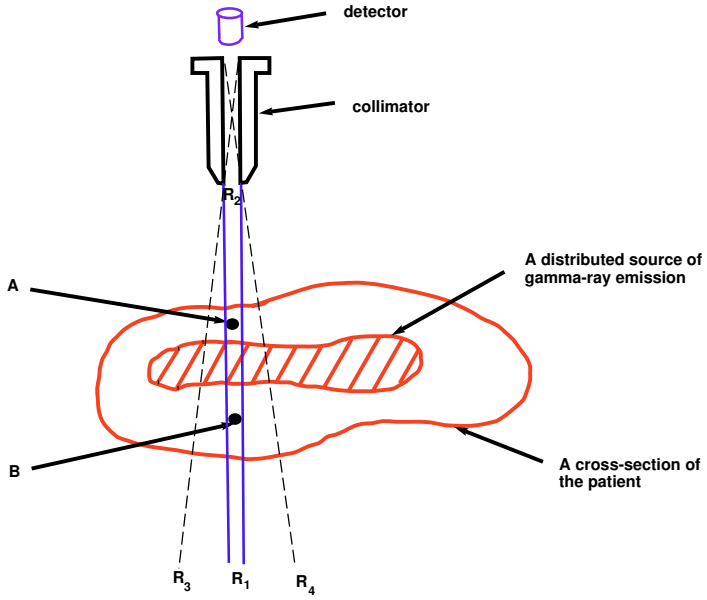


Figure 8.1: Single photon emission tomographic imaging.

A serious difficulty with tomographic imaging of a gamma-ray emitting source is caused by the attenuation that photons suffer during their travel from the emitting nuclei to the detector. The extent of this attenuation depends upon both the photon energy and the nature of the tissue. Consider two elemental sources of equal strength at points A and B in figure 8.1: because of attenuation the detector will find the source at A stronger than the one at B. A number of different approaches for attenuation compensation have been developed, but will not be discussed here.

8.1.3 Digital Filtering in SPECT

Filtered back-projection tomography (Kak and Slaney, 1988) (FBT) is commonly used for image reconstruction and was used in this study. Back-projection alone (without filtering) results in

[4]Strictly, this is the line integral of the photon flux along this path.

undesirable image smoothing and the presence of star-like artifacts in the reconstructed image. The degree to which back-projection artifacts can be removed must be balanced by the degree to which image noise can be tolerated. Frequency refers to the change in number of counts from pixel to pixel. True image signal falls off rapidly with increasing frequency, while the noise content remains constant. Background noise is considered to be of a high frequency because there is marked variability in the number of counts from pixel to pixel. Image sharpness (edges and fine detail) are also high frequency, while the target is low frequency.

Ramp filters (high pass filters) are used to boost high frequencies in order to sharpen the spatial details (edges) of the image. Unfortunately, this also increases the noise, because the filter linearly enhances higher frequencies (hence the term "ramp"). Thus, although ramp filters produce the highest resolution possible in a reconstruction, the images are often uninterpretable due to the propagation of noise associated with low count statistics. To limit this effect (*i.e.* to decrease the noise) a second roll-off or low pass (smoothing) filter is applied. A low pass frequency filter is employed to reduce noise. Low pass filters increase the signal to noise ratio, but at the expense of image contrast and resolution.

When selecting a processing filter there is always a trade-off between image contrast and image uniformity. The cut-off frequency is the frequency above which all data is removed. The lower the cut-off frequency employed, the smoother (more uniform) the reconstructed image and the greater the loss of contrast and resolution due to the loss of image sharpness contained in the higher frequency data. High count statistics are crucial, as the higher the number of counts in the projected data, the higher the cut-off frequency can be. A filter with a high cut-off produces images with a lot of contrast which can result in a high sensitivity (detectability of target), but low specificity (classifiability of the target). The filter order refers to the steepness of the slope of the filter curve. High order implies a steep slope. This produces a sharper image, but also creates more image distortions.

As just discussed, digital filtering of the data is an integral part of the reconstruction process in tomography. Although the choice of filter parameters has a dramatic effect on both visual quality and spatial accuracy of the data, the decision depends on many parameters and there is no easy, or even unique, answer. In non-quantitative SPECT, subjective tests, such as viewer preference, may be used to determine an "optimal" cutoff frequency which makes a trade-off between resolution and noise. For quantitative studies, it is important to have an objective algorithm to calculate the appropriate filter parameters. This takes us into the next section.

8.2 Development of a Measure of the Effects of Image Filtering

As mentioned in the last section, the quality of the final SPECT image depends on the values of parameters in the chain of transformations from patient to observer. This chain begins with the physical properties of the patient and isotope combination, and ends with the perceptual response of the observer. This chain has been referred to as the *object to observer pipeline* (OTOP) (Klymenko et al., 1990). Many of the parameters of the OTOP are relatively fixed,

except for the image processing stages, such as restoration filtering. There exists some freedom here to vary the filtering parameters, and these impact greatly on the final image quality. In order to optimise these parameters in terms of the final subjective quality of the SPECT image, it is necessary to define a metric that produces a satisfactory measure of the effects of the variations in the parameters of the filtering process on the image undergoing transformation. Therefore the main aim of the work described in this chapter was to develop an image metric that predicts the effects of different filtering functions and parameters when used on SPECT images.

8.2.1 A Gradient Measure

A measure of the effects of filtering on ECT and reconstructed[5] (TV) SPECT images was developed. This measure was based on an image quality measure developed for evaluating compressed video quality (Quincy, 1990), where the measure is derived from the square of the number of pixels which are the output of an edge detection filter and have values above a grey-level threshold. Quincy obtained this threshold by a tedious method of subjective scoring of image quality. The measure to be described in this chapter was developed to quantify the amount of blurring and noise in the SPECT images after filtering, and differs from Quincy's measure by introducing an objective way of obtaining the threshold. This then facilitates the optimisation of the filtering parameters in terms of image quality, as discussed elsewhere in this chapter.

The measure, which is derived in this chapter, is called the gradient energy measure (GEM) and is based upon the *Laplacian* and *Sobel* edge detecting convolutional filters (Gonzalez and Wintz, 1987b). These tend to emphasise the edges; the Sobel tends to operate on line edges, while the Laplacian tends to operate on point intensity discontinuities. A brief description of these filters is now given.

Gradient Operators

The filters used here are edge or gradient enhancers and are applied under the operation of convolution[6]. The convolution of a digital image f, with a filter function h, results in the filtered image g and is defined by

$$g(x,y) = \sum_{i=0}^{N-1} \sum_{k=0}^{N-1} h(x-i, y-k) f(i,k), \tag{8.1}$$

where $x = 0, 1, 2, \ldots, N-1$ and $y = 0, 1, 2, \ldots, N-1$, for an $N \times N$ array of filter coefficients h, called a *spatial convolutional mask* or just "mask". An example of a 3×3 mask is shown in figure 8.2, where N is usually odd.

[5]The reconstructed images used here are transverse sections (slices) of the brain and hence are called TV images.

[6]Convolution is discussed, in the context of point spread functions, in section 2.2.4 of Chapter 2.

w_{00}	w_{01}	w_{02}
w_{10}	w_{11}	w_{12}
w_{20}	w_{21}	w_{22}

Figure 8.2: A 3×3 spatial convolution mask.

The convolution defined in (8.1) for $N = 3$, is achieved conceptually by superimposing the array h over the input image f and then taking the average of the product of each term in the mask with its associated image pixel from f and replacing the pixel "under" w_{11} with this new value; *i.e.* the new pixel value $p = \frac{1}{9}\sum_{i=0}^{2}\sum_{j=0}^{2} w_{i,k} f i, k$, for $N = 3$.

This mask is then displaced by one pixel and a new value for this central pixel is calculated. This is continued until the each of the input image pixels has been recalculated[7]; *i.e.* has been the pixel "under" the mask central pixel.

We will now briefly consider the Laplacian and Sobel filters and their implementation as convolutional masks.

Sobel Operators

Consider an image $f(x, y)$, the gradient at the point (x, y) is defined as

$$\vec{G}[f(x,y)] = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{8.2}$$

It is well known from vector analysis that the vector $\vec{G}$ points in the direction of maximum rate of change of f at (x, y). For edge detection, we are usually only interested in the magnitude of this vector, which is known as the *gradient* and is defined as

$$G[f(x,y)] = [G_x^2 + G_y^2]^{\frac{1}{2}}. \tag{8.3}$$

For ease of implementation, this equation is usually approximated by

$$G[f(x,y)] \approx [|G_x| + |G_y|]. \tag{8.4}$$

The gradient defined in (8.2) can be implemented in a digital fashion in various ways. It is well known that derivatives can be approximated numerically by taking differences. This is commonly done in approximating (8.2), by taking a 3×3 mask, centred at (x, y), as follows. Consider the sub-image shown in figure 8.3(a), where x_{11} represents the grey-level location at (x, y), with the other x_{ij} representing its 8 neighbours. Then the component of the gradient in the x direction is defined as

$$G_x = (x_{20} + x_{21} + x_{22}) - (x_{00} + x_{01} + x_{02}), \tag{8.5}$$

[7] Actually, the pixels around the perimeter of the input image, to a depth of (next highest integer of $\frac{N}{2}$) $- 1$, are untouched; *e.g.*, depth equals 2 pixels for $N = 3$

x_{00}	x_{01}	x_{02}
x_{10}	x_{11}	x_{12}
x_{20}	x_{21}	x_{22}

(a)

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

(b)

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

(c)

Figure 8.3: The Sobel operators: (a) 3×3 image region, (b) Mask for G_x, (c) Mask for G_y.

and in the y direction as

$$G_y = (x_{02} + x_{12} + x_{22}) - (x_{00} + x_{10} + x_{20}). \tag{8.6}$$

The masks shown in figure 8.3(b) and in figure 8.3(c), which are known as Sobel operators, perform the operations in (8.5) and (8.6) respectively. The outputs of these two masks are combined, using (8.4) to obtain the gradient at (x, y).

Laplacian Operator

The Laplacian is a second-order derivative operator defined as

$$L[f(x,y)] = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}. \tag{8.7}$$

Using differences again to approximate differentials in digital images, the mask shown in figure 8.4 implements (8.7) digitally at the point (x, y) with reference to the region of the input image f, shown in figure 8.3(a), as defined before. When convolved with the image, this mask computes the digital Laplacian, L_{xy}, by performing the operation

$$L_{xy} = x_{10} + x_{21} + x_{12} + x_{01} - 4x_{11}. \tag{8.8}$$

This functions as a second-order derivative, in that the output of the operation is zero in constant image areas or when the region of support of the mask is on the ramp of an edge. Being a second-order derivative, the Laplacian is quite sensitive to noise, as will be demonstrated later.

0	1	0
1	-4	1
0	1	0

Figure 8.4: Mask used to apply the Laplacian.

Now that the basic underpinnings of the GEM have been discussed, the development of this measure will now be detailed.

8.2.2 Development of the GEM

The combination of an edge filter with a threshold gives a measure of the amount of blurring in an image. This blurring is the result of the convolution of the image with an imaging system point-spread-function, which increases the amount of correlation between neighbouring pixels. Therefore, the level of the intensity discontinuities in the image are reduced, resulting in a lower number of pixels above a specified threshold in the edge filtered image. Another reason for using these filters is that the human visual system is very sensitive to edges and the Laplacian has been used as an approximate model of ganglion retinal cell effects (Young, 1986; Kingdom and Moulden, 1989). A measure M was applied to the images transformed by the Laplacian and Sobel filters. The measure M_j for the j^{th} image of an image set is:

$$M_j = \frac{1}{n}\sum_i g_i^2, \;\; g_i = \begin{cases} g_i, & g_i > T \\ 0, & \text{otherwise} \end{cases} \tag{8.9}$$

where g_i is the value of the i^{th} pixel above the threshold T and n is the number of pixels in the image with a value above T.

The measure M for an image set or subset is then just the mean value for M_j over the set:

$$M = \frac{1}{N}\sum_j^N M_j \tag{8.10}$$

Thus M is a measure analogous to the average pixel energy, since it represents the average squared value of the pixels greater than the threshold T, and has units of grey-level squared per pixel, but since grey-level and pixel are unitless, so is M.

The measure defined in equation (8.9) was chosen as it was expected to have certain properties. For a fixed threshold, the GEM was expected to increase monotonically with image noise power and decrease monotonically with increase in image blurring. Also, the GEM was expected, as a function of threshold, to have a single maximum; *i.e.* coincident local and global maxima. This behaviour is demonstrated in figure 8.5, with the underlying components of the measure, which sum to give the expected behaviour. Thus, the threshold value would be obtainable automatically from the data.

8.3 Experimental Evaluation of the Measure

This section provides the combined methods and results for a series of experiments evaluating the efficacy of the GEM in measuring the effects of various kinds of filtering on SPECT images. The filtering types that are considered are: Butterworth low pass (LP) filtering, filtering by convolution of an image with Gaussian point spread functions (PSF), Weiner filtering, and the addition of Poisson noise. Images were produced using each of the listed filtering processes in separate sets. The GEM was calculated for each image set using equations (8.9) and (8.10), where N in the latter was equal to 10; *i.e.* each set consisted of 10 images. Note, each set of images were gathered from a single acquisition from the same patient.

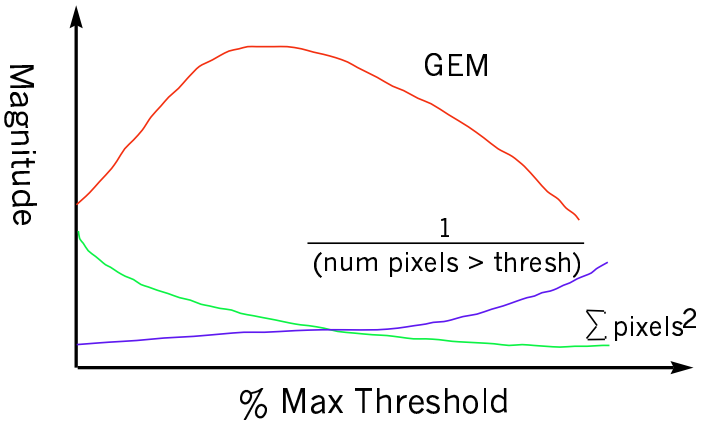


Figure 8.5: Expected behaviour of gradient measure as a function of threshold.

The experiments were performed by plotting the GEM output for each filtering type at various levels of the respective filtering parameters for each filtering type. These plots also summarise the experimental configurations that were explored, including the filtering operations and the levels of the filter parameters that were used. Plotting was found to be sufficient for analysis since the GEM output level was not stochastic with respect to the filter parameter level; *i.e.* for a given set of images, there was no variation in optimum threshold value or GEM output for a given filtering operation. However, limited subjective experiments were done, which required simple statistical analysis.

8.3.1 Subjective Analysis

Some limited experiments on the subjective quality of filtered images were done to compare with the GEM performance, to give an indication as to its suitability as a predictor of subjective image quality.

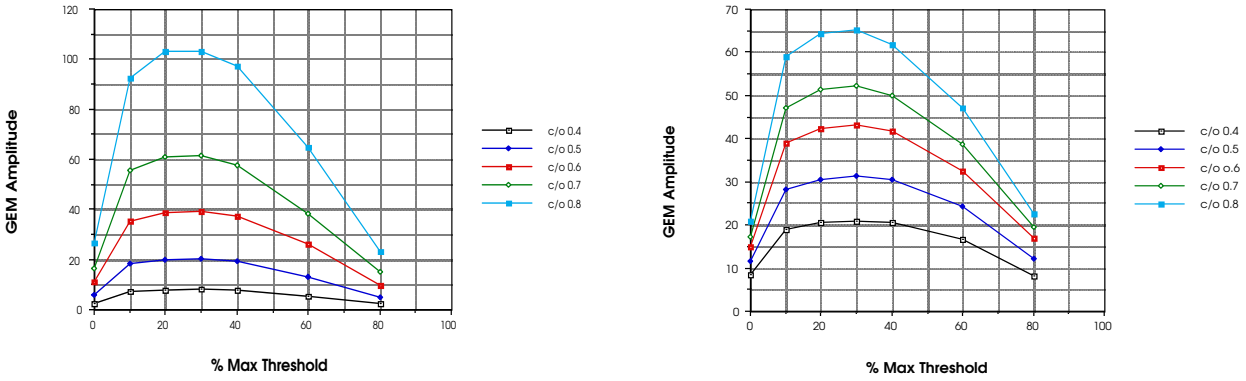
In these subjective studies 4 expert users (nuclear medicine physicians) were used. They were required to rate, on a scale from 1 to 10, the effect of filtering on each image to which the GEM was applied. They were told to rate: sharpness of the image for low pass and PSF filtering, noise level for images with added noise, and both sharpness and noise level for Weiner filtered images. The images were presented on Sun workstations under their usual working conditions; *i.e.* standard office conditions.

8.3.2 Measurement of Low Pass filtering Effects

As detailed later many experiments were performed to evaluate the behaviour of the measure. Initially, highly controlled data, from the digitised Hoffman phantom[8] data set, were used. These data were low pass filtered, using a Butterworth filter, with cutoffs (c/o) in the range 0.4 - 0.8

[8]This is a brain phantom, or test object, with construction based on an accurate imaging of a human (Hoffman's) brain.

cyc/cm, producing several images. The GEM was applied to these images, with the threshold being varied from 0% to 80% of the maximum pixel value for each image. The results are shown in figure 8.6(a), for the Laplacian, and in figure 8.6(b), for the Sobel . It was found that the optimum threshold value, which obtained the peak output of the GEM, was robust with respect to the amount of LP filtering, although the actual curve became flatter as the blur increased; *i.e.* the higher the amount of blur, the more insensitive the GEM was to the threshold value. This behaviour was consistent with expectations.



(a) Using the Laplacian operator.

(b) Using the Sobel operator.

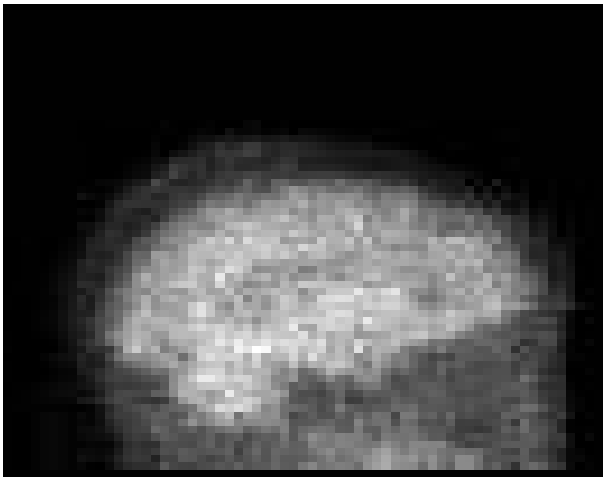
Figure 8.6: Output of the GEM for Hoffman digital data versus the threshold (T) value, and for different cutoff (c/o) frequencies of the low pass Butterworth filter.

Examples of actually acquired[9] and Butterworth filtered Hoffman phantom images are shown in figure 8.7, for ECT images, and in figure 8.8, for TV images. To test the GEM, as a measure of smoothing (the cutoff frequency) due to low pass filtering, the GEM amplitude was plotted against the Butterworth filter cutoff frequency for both digitised and acquired Hoffman image sets. This was done with the threshold set to the optimum value; *i.e.* where the gradient of the GEM versus % threshold curve is equal to zero. Also plotted on the same graph is a subjective score given to the images in terms of perceived smoothness. This is shown in figures 8.9(e) and 8.9(f), where it can be seen that the GEM behaves as a monotonically increasing function of the cutoff frequency with all the image sets. All the curves, including the subjective graph, appear to be strongly correlated. This is demonstrated in table 8.1 which lists a correlation analysis for the data plotted in figures 8.9(e) and 8.9(f), and shows very high correlation between the curves.

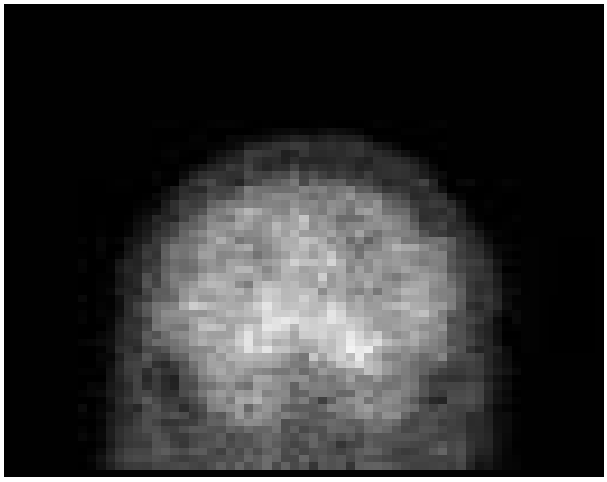
8.3.3 Measurement of Noise and Point Spread Function Effects

The GEM has been shown to be useful for measuring blur due to LP filtering. However, a question arises concerning blurring due to PSF effects and the effects of noise. Accordingly, a series of experiments were performed to determine these effects. Initially, a reconstructed (TV)

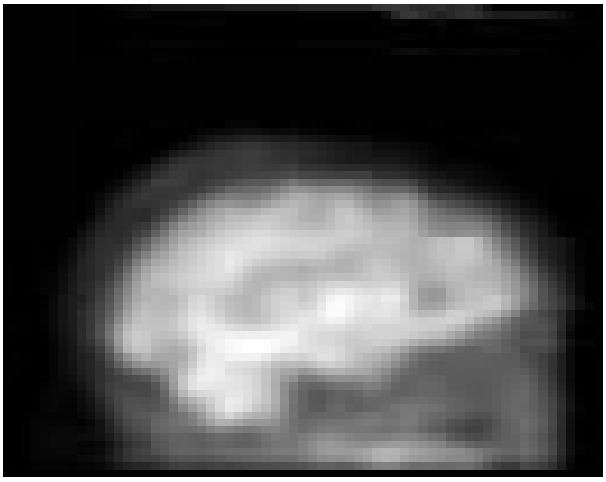
[9]Images produced by scanning of a Hoffman phantom filled with radio-isotope, usually Technetium 99.



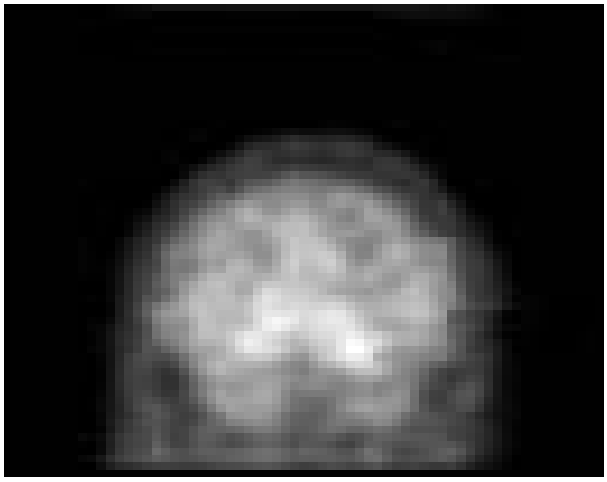
(a) Raw ECT image - Lateral view



(b) Raw ECT image - posterior view



(c) Butterworth filtered (0.5 cyc/cm) ECT image - Lateral view

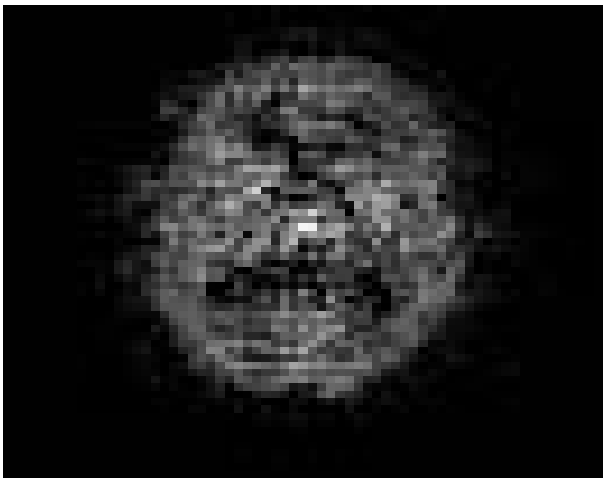


(d) Butterworth filtered (0.7 cyc/cm) ECT image - posterior view

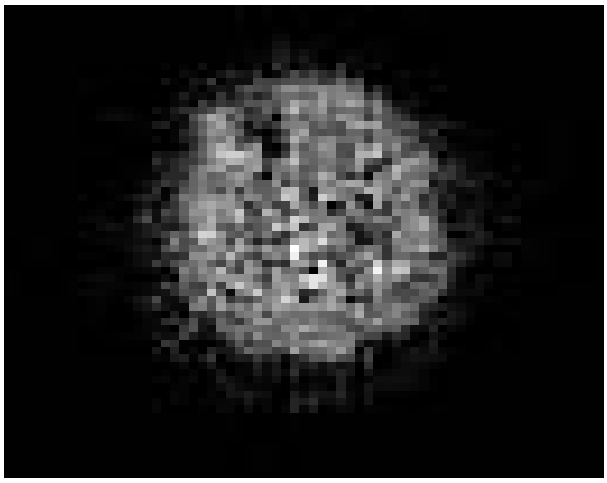
Figure 8.7: Low pass filtered and unfiltered emission (ECT) images.

Pearson Product-Moment Correlation							
	lap ect	lap tv	sob ect	sob tv	lap dm	sob dm	subjective
lap ect	1.000						
lap tv	0.996	1.000					
sob ect	0.963	0.981	1.000				
sob tv	0.992	0.979	0.924	1.000			
lap dm	0.993	0.983	0.934	0.992	1.000		
sob dm	0.992	0.999	0.987	0.970	0.978	1.000	
subjective	0.998	0.997	0.973	0.986	0.983	0.993	1.000

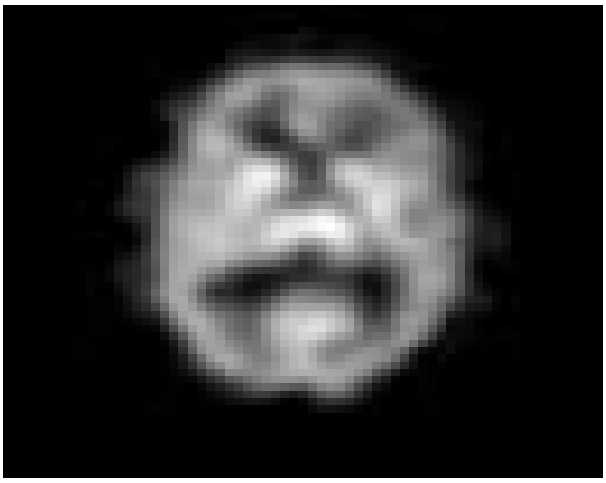
Table 8.1: Correlation of GEM for Hoffman digitised, ECT & TV and subjective scores.



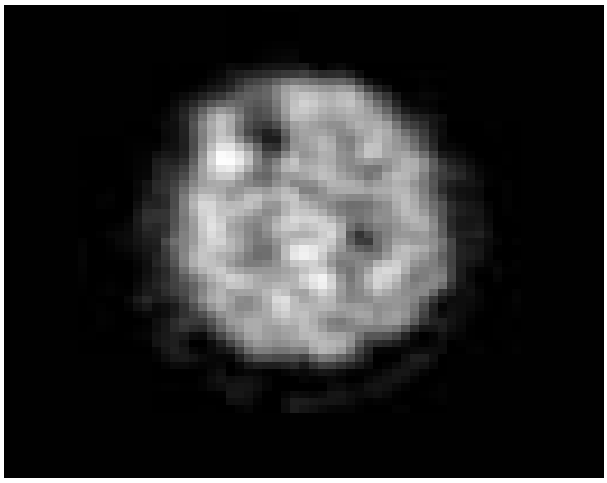
(a) Raw TV image - Lateral view



(b) Raw TV image - posterior view

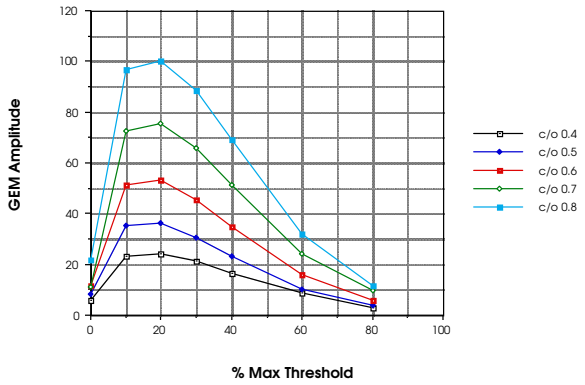


(c) Butterworth filtered (0.5 cyc/cm) TV image - Lateral view

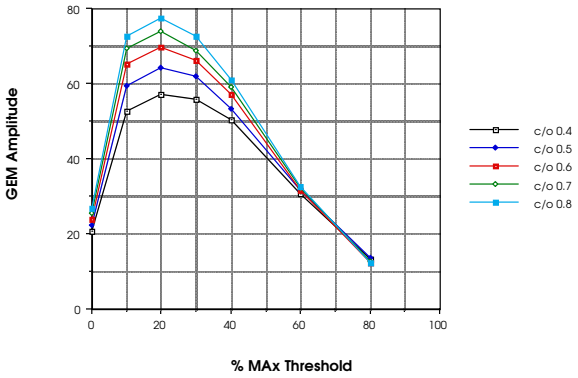


(d) Butterworth filtered (0.7 cyc/cm) TV image - posterior view

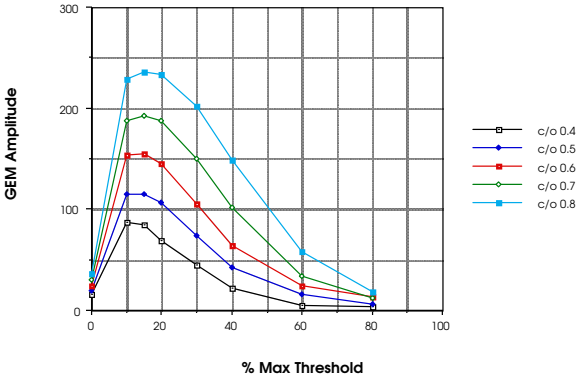
Figure 8.8: Low pass filtered and unfiltered reconstructed transverse (TV) images.



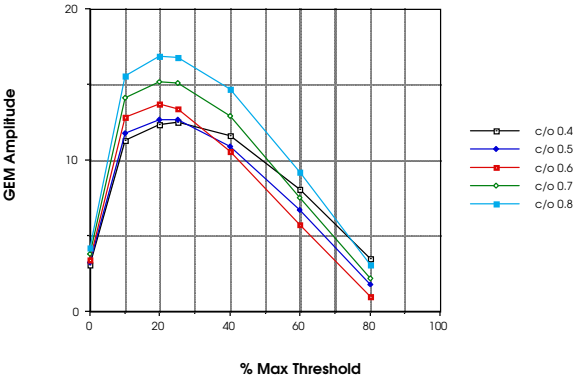
(a) GEM with Laplacian on ECT images.



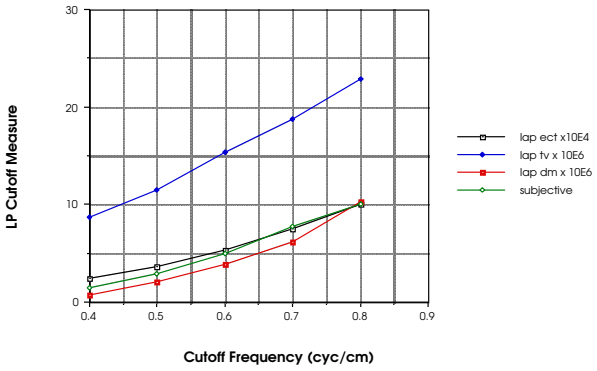
(b) GEM with Sobel on ECT images.



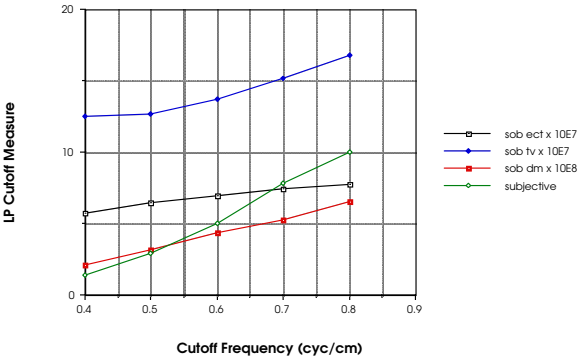
(c) GEM with Laplacian on TV images.



(d) GEM with Sobel on TV images.

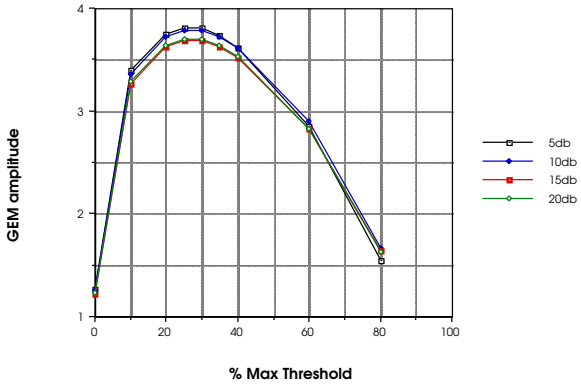


(e) LP smoothing measure (Laplacian).

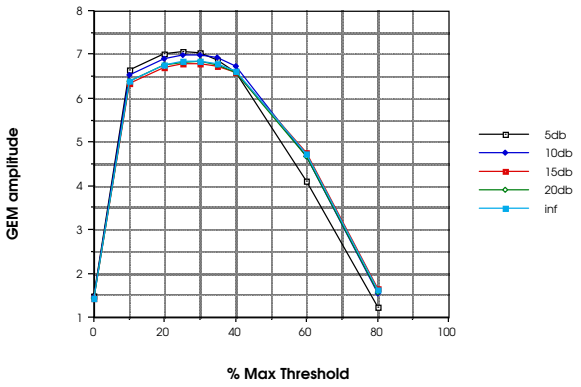


(f) LP smoothing measure (Sobel)

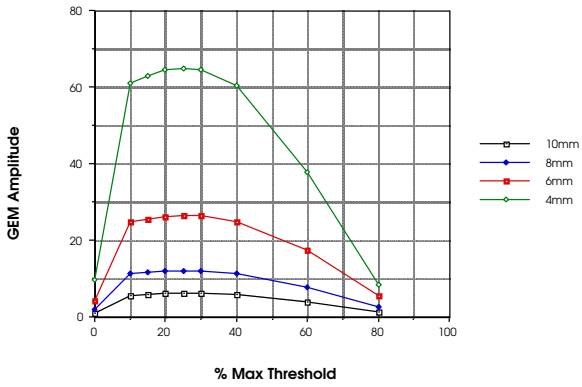
Figure 8.9: GEM as a measure of low pass (LP) cutoff (c/o) frequency. Sub-figures (a) to (d) plot GEM output vs threshold for various LP c/o frequencies, while sub-figures (e) & (f) plot GEM output as a function of c/o frequency with the threshold constant at the optimum value. ECT = emision computed tomography, TV = Trans Verse (reconstructed)



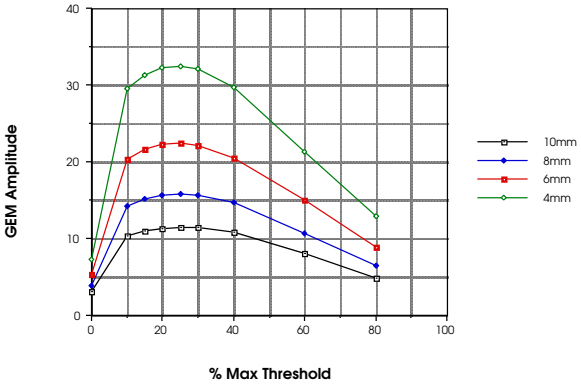
(a) Noise, Laplacian operator.



(b) Noise, the Sobel operator.

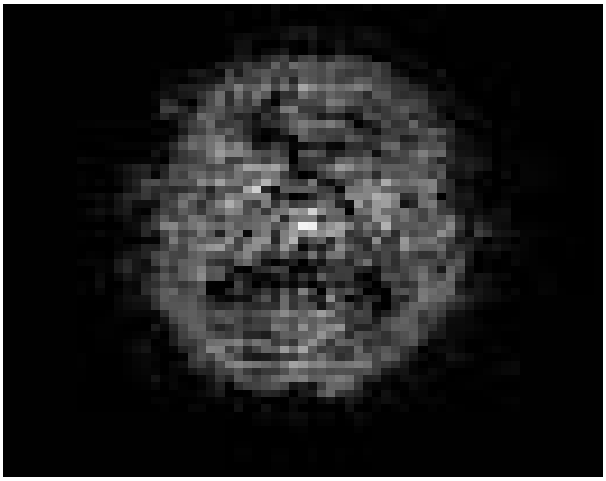


(c) PSF FWHM, Laplacian operator.

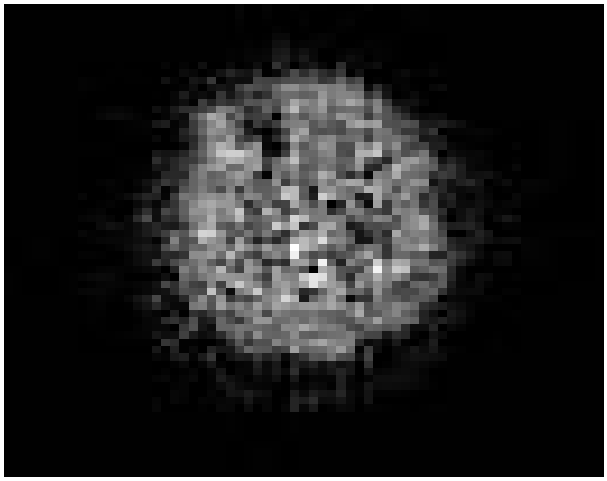


(d) PSF FWHM, Sobel operator.

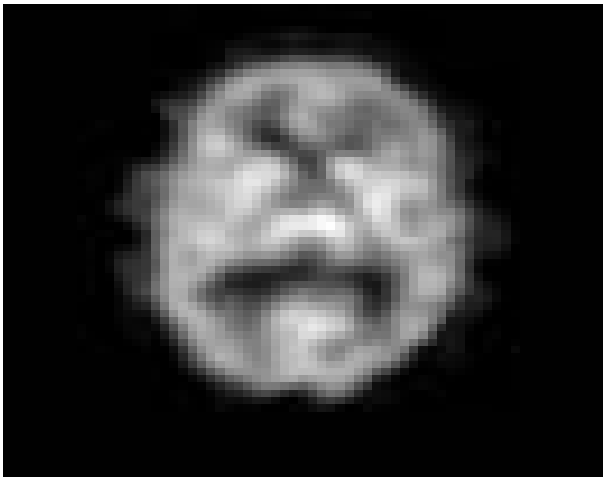
Figure 8.10: GEM threshold robustness. Shown is GEM output vs threshold for various levels of noise and PSF full-width-half-maximum (FWHM) in millimetres (mm).



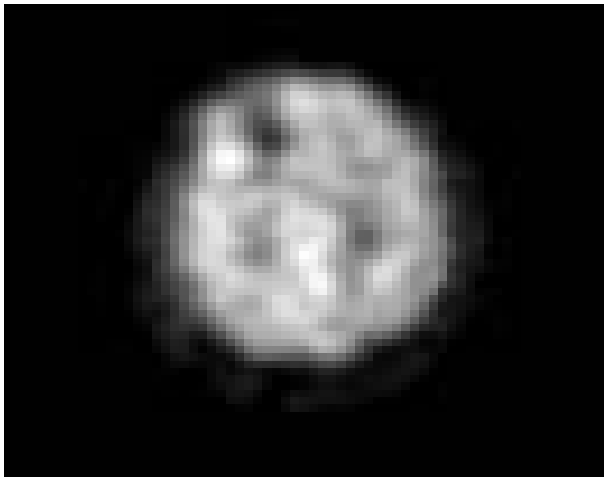
(a) Raw TV image - Lateral view



(b) Raw TV image - posterior view



(c) Convolved with PSF (60mm FWHM) TV image - Lateral view



(d) Convolved with PSF (60mm FWHM) TV image - posterior view

Figure 8.11: Reconstructed images convolved with Gaussian PSF.

acquired Hoffman image set was convolved with a PSF from a constructed database of Gaussian PSF convolutional kernels. Different versions of this image set were then created by adding a different level of Poisson noise to each set. The noise level was defined in terms of the signal-to-noise ratio (SNR), specified in decibels[10] (db). It was found that the value for the optimum threshold value for the GEM was insensitive to noise (see figures 8.10(a) and 8.10(b)). However, even though the GEM threshold value was robust with respect to noise, the actual output of the measure was a monotonic function of the noise (see 8.12(a)), showing it was also a noise measure. As with LP blur, experiments were performed to ascertain the effects of Gaussian PSFs with varying *full width half maximum*[11] (FWHM)s on the stability of the optimum threshold. Figure 8.11 shows an example of blurring due to convolution of a reconstructed image with a Gaussian PSF. The behaviour of the GEM was found to be similar to that of low pass filtering, and is shown in figures 8.10(c) and 8.10(d).

The value of the GEM, at optimum the threshold of 25% (as seen in figures 8.10(c) & 8.10(d)), was plotted against FWHM (see figure 8.12(b)), and was found to behave well as a measure of PSF-induced blur. The Laplacian produced a larger dynamic range than the Sobel based measure. Also plotted is a subjective score of perceived image blurring. This correlated well (see table 8.2) with both versions of the GEM, but showed slightly better agreement with the Sobel based measure.

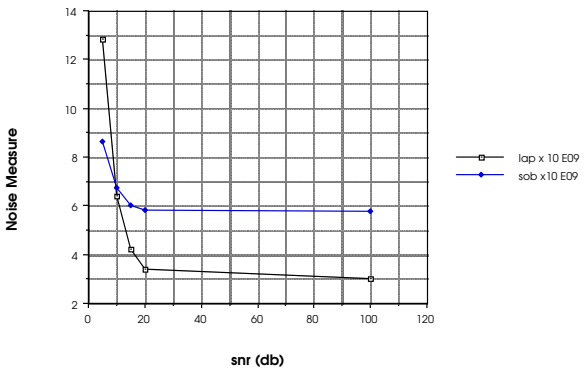
Pearson Product-Moment Correlation			
	sob x 10^8	lap x 10^7	subjective
sob x 10^8	1.000		
lap x 10^7	0.982	1.000	
subjective	0.999	0.972	1.000

Table 8.2: Correlation of GEM for Laplacian and Sobel measure with a subjective score for Gaussian blur.

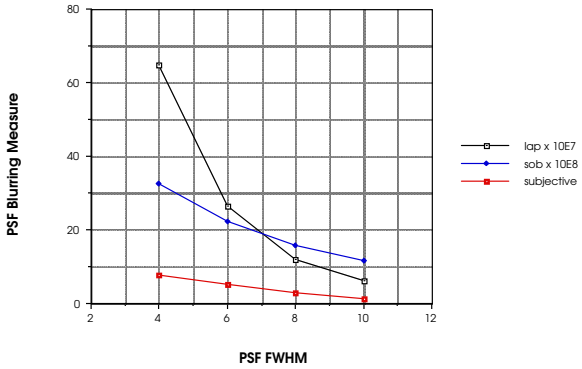
Further analysis was carried out to investigate the combined effects of noise and convolutional blurring. The results are summarised in figures 8.12(c) and 8.12(d), which show the logarithm of the GEM amplitude versus the noise level in db (down from the signal level), for various widths of the PSF. These graphs show that the GEM amplitude is a monotonic function of noise level at a particular PSF width, and for any given abscissa value (within the domain of all the functions plotted) the ordinate values for plots at decreasing PSF widths, are monotonically increasing. There is also evident an interaction between noise level and PSF width, as the gradient of each curve becomes steeper as the PSF width decreases. This is consistent with commonly known facts about smoothing operations (convolution of image and PSF) and noise levels in images.

This indicates that the GEM is also a potentially useful measure of PSF induced blurring and noise level, where both these processes are occurring simultaneously in the images under consideration.

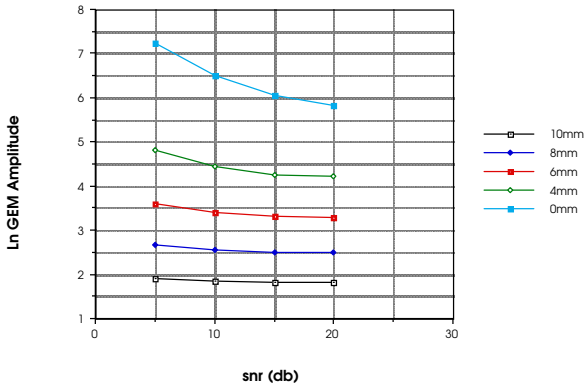
[10] SNR in decibels $SNR_{db} = 10 \log \frac{S}{N}$, where S is the image data and N is the noise level.
[11] FWHM is the width of the PSF where the amplitude has dropped to half its peak value.



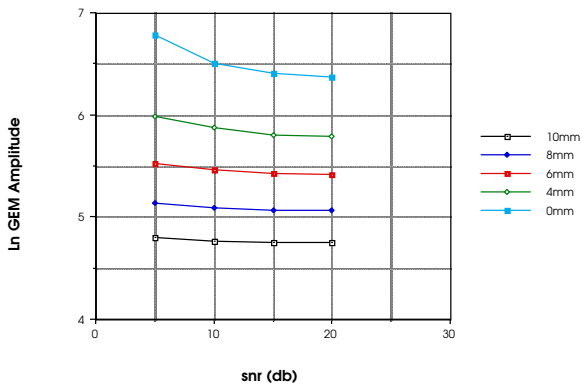
(a) As a noise measure.



(b) A measure of Gaussian blurring.



(c) A measure of both noise and PSF blurring (Laplacian).



(d) A measure of both noise and PSF blurring (Sobel).

Figure 8.12: GEM as a measure of noise and/or PSF blurring.

8.3.4 Measurement of Weiner Filtering Effects

An important aspect of SPECT image processing is that of restoration filtering, which is an attempt to gain an optimum tradeoff between noise and blurring in the final image. A filter that is commonly employed in this function is the Wiener filter (Gonzalez and Wintz, 1987c). An example of a Wiener filtered emission image is shown in figure 8.13. A version of this filter (a parametric Wiener filter) is characterised by the following equation for its power spectrum:

$$W(f) = m_{tf}(f)^{-1} \frac{m_{tf}(f)^2}{m_{tf}^2(f) + \gamma \frac{N^2(f)}{B^2(f)}}. \tag{8.11}$$

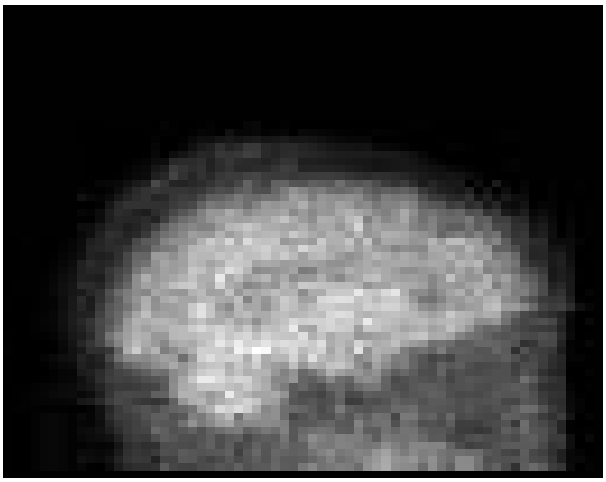
Where m_{tf} is (ideally) the equivalent modulation transfer function of the imaging system; N^2 is the noise power spectrum; B^2 is the power spectrum of the object of interest, while γ is the free parameter which varies the tradeoff between noise and blurring.

The spectrum of a degraded image can be modelled by equation 8.12, where P_g is the degraded spectrum and P_f is the undegraded image power spectrum. In image restoration, the aim is to obtain P_f from P_g.

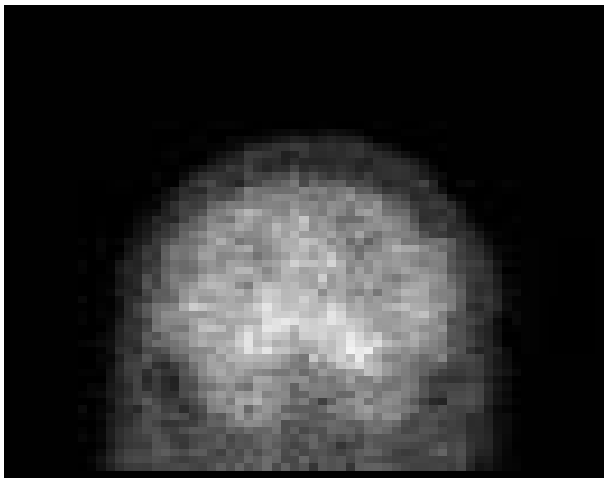
$$P_g(u,v) = N(u,v) + m_{tf}^2(u,v) \cdot P_f(u,v) \tag{8.12}$$

Experiments were performed to evaluate the GEM as a measure of the effects of the free parameter γ. A Wiener filter was applied to the acquired Hoffman ECT image set, with γ varying between 0.5 and 5.0 in 0.5 steps. The GEM was applied to both the ECT and the resultant reconstructed image sets. The results are summarised in figures 8.14 and 8.15. Figure 8.14 shows monotonically decreasing curves for increasing gamma, using both Laplacian and Sobel filters. Thus the GEM appears to be a good measure of the effects of gamma.

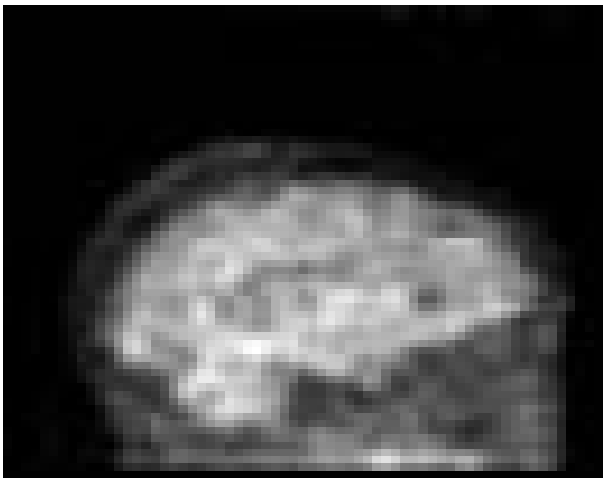
Figure 8.15 shows a scattergram, produced from the values of the GEM obtained for the ECT image set, versus the GEM values obtained for the image set reconstructed from this ECT image set. The regression line produced an R^2 value of 0.995, which suggests a very high correlation of the two sets of GEM values. This implies that there exists a linear relationship between the blurring in the ECT images and the blurring in the final reconstructed images.



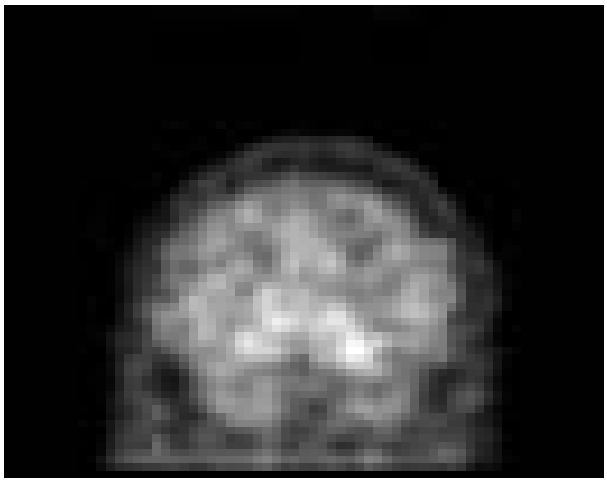
(a) Raw ECT image - Lateral view



(b) Raw ECT image - posterior view



(c) Wiener filtered ECT image - Lateral view



(d) Wiener filtered ECT image - posterior view

Figure 8.13: Wiener filtered emission images.

Figure 8.14: GEM as a measure of γ in Weiner filtering.

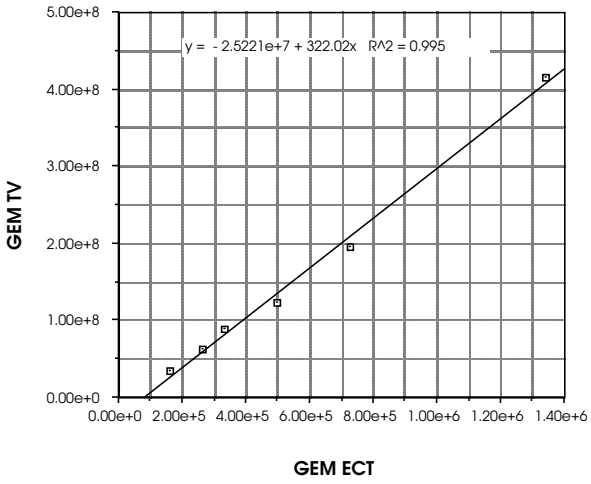


Figure 8.15: Scattergram & regression line of GEM for ECT vs reconstructed image set.

8.4 Conclusions and Further Work

The gradient energy measure was demonstrated to be a useful measure of the effects of several commonly used filtering operations on SPECT images, and was sensitive to the filtering parameter settings. It was also found to agree well with preliminary subjective ratings. Laplacian and Sobel versions worked well, though the former correlated slightly better with subjective data. The Sobel filter was more insensitive to noise, which may make it more suitable in some situations.

The fact that the GEM correlates well with subjective ratings means that it could possibly be used as a stand-alone measure, though this is not necessary for its intended use in optimising SPECT image quality using full-blown subjective experiments.

There is a model of HVS function that appears to be modelled to a first order by the GEM. This model is known as the "energy integrator" model (Green and Swets, 1966c; Moulden et al., 1990). To determine if this is an edequate model, more work needs to be done.

It is appropriate that this measure be used in conjunction with subjective analysis of SPECT images in order to find the optimum filter parameters in terms of clinical image quality. This will require a large subjective study to ascertain the optimum set of processing parameters and to calibrate the GEM. Methods of subjective image quality analysis are discussed in chapter 3 and it is recommended that the receiver operating characteristic (ROC) method is the most appropriate, though some insight, where subjective confidence is not a concern, could be obtained from using an analysis-of-variance (ANOVA).

Chapter 9

The Effects of Clutter on Human Target Detection Performance

Summary: *This chapter describes a study which determined the performance of human image analysts in the surveillance context, using Synthetic Aperture Radar (SAR) derived images, in terms of the analyst's receiver operating characteristic. The experiment was designed to correspond as closely as possible to the expected real world mode of operation of the analysts using similar imagery. In particular, the effects of target contrast and background clutter on human analyst target detection performance were quantified.*

9.1 Introduction

This chapter contains a description of a major study to determine the performance of human image analysts, using a synthetic aperture radar (SAR) (Fitzsimons, 1998; Oliver and Quegan, 1998) imaging system, which is currently under development. The study was designed to obtain practical baseline data on human analyst performance, and to investigate the issues of clutter and contrast measurement.

It is likely that a synthetic aperture radar system will be used to detect possible targets of interest in remote regions of Australia's north. This may be of significance to customs, police, search and rescue or military operations.

There are three main requirements for operational effectiveness of a SAR system in this type of context. Firstly, the information it gathers must be near real-time. Secondly, the probability of detection for any target of interest in the region of operation must be high. Thirdly, the false alarm rate of the SAR system must be low enough to ensure that the limited resources used to follow up each SAR detection are not wasted.

The issues of timeliness and the probability of detection and false alarm rate (FAR) for the entire SAR system will not be addressed here. This study concentrates on measuring the performance, in terms of probability of detection and false alarm rate (called the *receiver oper-*

ating characteristic (ROC)), of one component of the SAR system — the human analyst. In the operational system, a number of analysts will be employed to examine the SAR imagery in order to detect and report any possible targets of interest. As a result, the overall performance of the SAR system depends initially upon the performance of human analysts in this visual detection task.

The study presented here determined the unassisted performance of human analysts in the SAR context in terms of the analyst's ROC. The experiment was designed to correspond as closely as possible to the expected (fully manual) *search-mode* of operation of SAR analysts using similar SAR imagery. In this mode, the analysts manually scanned through all the imagery looking for the possible targets of interest; *i.e.* no automatic target detection algorithms or other cueing methods were employed. Therefore, the tools used to view the image in the experiment closely matched the analyst's expected method of manipulation and display. In this way, the external validity of the experiment was made as high as possible (although for the sake of internal consistency, and to keep the size of the experiment down to a manageable level, certain limitations – specified later – were placed upon the analysts).

This experiment was designed to determine the search-mode analyst performance in order to arrive at a base-line, against which to measure the impact of automatic target detection on the whole system. When automatic target detection is employed, the analysts are only cued with cut-outs[1] of small regions from the imagery that the automatic target detection algorithms identify as possibly containing targets, thereby eliminating any search component from the analyst's work. This is termed the *cueing* mode. This change in methodology will impact significantly upon the performance of analysts (both in terms of their coverage rate and ROC) and is expected to improve it over the performance determined here.

This search-mode analyst ROC experiment was performed with targets inserted in imagery across a wide range of visual clutter and contrast levels. This does not reflect statistically the operational distribution of targets and clutter, and it was not meant to do so. It is important to emphasise that additional information is required to interpret these results in terms of the ROC performance of the SAR system with fully manual analysis in an operational scenario. In fact, several different types of information are required. They include: the distribution of foliage and terrain types across a region of operation, the distribution of targets, the imaging characteristics (including imaging geometry) of the radar, and the target-to-background contrast that occurs, given the target type, foliage, and imaging characteristics. This work is currently being done by others, and is beyond the scope of this thesis. However, these crucial pieces of information must be added to the results of this study to determine the ROC performance of the SAR system with fully manual analysis.

Ideally, this experiment should have been performed using imagery that had been exhaustively ground-truthed. However, no such imagery was available. Consequently, certain compromises were necessary in the experimental design. As a result, it was difficult to relate the analyst's performance to the SAR system's false alarm rate. In other words, ground-truthed

[1] Excised regions.

imagery would have allowed the quantification of the analyst's (search-mode) false alarm rate for actual targets of interest on the ground. Instead, what is reported here is their false alarm rate for target-like configurations of pixels, whether or not these were due to targets of interest. These issues are discussed in more detail in section 9.2.7.

9.1.1 The Visual Task

In the literature, ROC analysis is usually applied to a pure detection task; *i.e.* the signal is either present or it is not. In general, the stimuli are either noise or noise plus signal. In this case, the task is slightly more complicated. The experiment could have been set up as either:

(i) A discriminatory detection task, where a target has to be distinguished from the background or clutter;

(ii) A classification task, where potential targets have to be distinguished from non-targets that are similar in terms of radar cross section; *e.g.*, vehicles versus power poles.

In the classical discriminatory detection problem, the target may or may not be present in the clutter. In contrast, the classical classification problem presents the observers with one object from a number of classes of objects, and the observers have to indicate which class they think it is from. The experiment was designed for task (i) because it most closely matches the SAR analysis task. However, there is a classification component to the SAR analysis task in that once an analyst has detected a target-like object in the imagery they would normally try to distinguish it from cultural clutter, such as farm sheds and power poles, using contextual information. I have not measured these aspects of the problem. Because of the widely varying approaches of the observers (see Appendix K) no attempt was made to investigate these aspects of the problem, as a result, the experimental task is a purely discriminatory detection one. This trade-off is discussed in more detail in section 9.2.7.

9.2 Experimental Methods

This section details the experimental design. In this study, both analysis of variance (ANOVA) and ROC methods have been used (see chapter 3). The ROC analysis employed the rating method (section 3.2.3 on page 74).

The difference between the two analysis techniques is that ANOVA focuses on obtaining purely sensory information, which in contrast to the ROC analysis, does not give any information about the observer's internal criterion. ANOVA also gives much more information about the relationships between the factors and their effects, including, for example, interaction analysis. However, in the case of discriminatory detection that is used here, it is necessary to take into consideration the observer's criterion. Therefore ROC analysis, which gives a summary measure of both sensory and subjective criterion information, is also required.

9.2.1 Experimental Design

The experiment was designed to have fixed effects, with blocking, resulting in nine treatments. The power and significance estimation was based on a two-way analysis-of-variance.

The nine treatments, that constituted a 3×3 cell, were a combination of three levels of clutter and three levels of contrast (c) times target area product ca_t, which were found to be within the approximate perceptual range of c for a target of approximately 5×5 pixels.

I have followed the usual practice, in visual perception studies, of defining the (Weber) contrast as

$$c = \frac{\mu_t - \mu_b}{\mu_b} \tag{9.1}$$

where μ_t is the mean target luminance and μ_b is the mean local background luminance.

The target contrast and the value of local clutter metrics were calculated over a 64×64 pixel region with the target at its centre. This value of background region was chosen to correspond to about 3° field-of-view when the observer is fixated on the target, at a viewing distance of 50 cm. From the studies performed in Chapter 7 (see summary information in figure 7.6) this seems to be an appropriate size of region around the target to define a region of "local" clutter.

It is also important to note that, for small or low contrast targets, such as are typically found in SAR imagery, the luminance threshold is a function of target area (a_t) (Blackwell, 1946; Lukis and Budrikis, 1982). To keep the experiment a manageable size, rather than use c as an independent variable, the ca_t product was used to control the detectability of the target. In this way, two physical target parameters (target contrast and target area) were being manipulated, but only their product was considered an independent variable in the experiment. This allowed a more representative set of targets in the experiment. On the other hand, as the responses were collapsed across target area and target contrast for each level of ca_t product, no specific information was available about the effect of target area or target contrast alone.

For this study the clutter levels were determined by applying a clutter metric due to Waldman *et al.* (1988). This metric has not been rigorously evaluated in the literature with respect to its correlation with subjective response, but which showed great promise in pilot studies. This clutter metric is discussed in detail in section 2.6 of Chapter 2.

9.2.2 Experimental Parameters

The statistical parameters for this experiment, which are defined in Appendix C, are now given. The significance was set at the 95% level ($\alpha = 0.05$) and the desired power was set at 90% ($\beta = 0.1$). The effect size that can be detected by the experiment is quantified by the parameter d; for this experimental design $d = 0.25$. This choice of parameters means that a small to medium effect should be detectable by the experiment. By use of a set of power tables (Cohen, 1977) this value of d corresponds to $f = 0.1$, under the assumption that the effects of the treatments are distributed uniformly along the range of spread; this statistic is tested in the F-test.

For this parameter set, the number of observations required was 240 per treatment. These observations were spread across each of 10 observers who saw the same set of (blocked) stimuli. This corresponds to each observer viewing 24 stimuli at each treatment level. For the sake of estimating the number of targets to insert, it was assumed that half the stimuli would contain targets (See the next sub-section 9.2.3 for further discussion). This corresponded to 12 inserted targets per observer at each treatment. All observers viewed the same imagery. Given that there were 9 treatments, the 12 inserted targets per observer per treatment level corresponded to a requirement for 108 targets to be inserted in total.

Ideally, each observer should be presented with the same set of stimuli, but in a different random order. However, for the sake of external validity, this was not possible in this experiment for two reasons. Firstly, observer's had complete control of the order in which they searched through the imagery. Secondly, there was a spatial relationship between the stimuli because they were embedded into a swath. Accordingly, all observers viewed the swaths in the same order.

9.2.3 Preparation for the ROC Analysis

In the literature on ROC curve studies, it is usual to set the *a priori* probability of detection to about 0.5; *i.e.* half the stimuli contain targets. Usually in ROC experiments, stimuli are presented to observers one after the other, or together, in a controlled way. This approach makes it easy to set up the required *a priori* detection probability. In the case of the SAR image analysis, the practice is more complicated. Here, image analysts examine large swaths of image data, which can be viewed by scrolling and panning, or by stepping through a screen-ful at a time. For the sake of external validity it was important that this operating procedure also be employed in the experiment.

To stay as close as possible to the normal operating procedures of analysts, yet at the same time accommodate the requirements of ROC analysis, the observers were required to examine a swath (or portion of a swath) one screen-ful (a frame) at a time, panning across the swath in a non-overlapping manner. In order to set up the ideal *a priori* detection probability of 0.5, the chance of a frame containing a target must equal the chance of a frame not having a target. This could have been obtained by dividing the swath into frames, with half of the frames containing a target. However, this was not possible in this experiment because it would have resulted in an atypically high target density. It is expected that the high target density would have modified observer behaviour in an undesirable way; *i.e.* the observers would have most likely learnt the probability of finding a target and thus anticipated a target detection, thereby distorting the results.

To circumvent this problem, a method known as the free-response ROC (FROC) (Bunch et al., 1978) was used. The FROC is applicable to situations that involve any number of reported locations and any number of actual targets in each image. Usually in the FROC, the abscissa represents the average number of false alarms per image presentation, whereas the ordinate is

Image	Resolution (m)	Pixel count	Equivalent Area[a] (km^2)	Target count
01	2	90947584	91	0
02	2	90947584	91	4
03	2	146857984	147	7
04	2	146857984	147	7
05	2	161021952	161	6
06	2	161021952	161	7
07	2	161021952	161	5
08	2	160276480	160	4
09	2	160276480	160	4
10	2	160276480	160	1
11	2	158785536	159	5
12	2	158785536	159	2
13	2	158785536	159	5
14	2	158040064	158	4
15	2	158040064	158	4
16	2	158040064	158	4
17	1	133218304	133	6
18	1	160145408	160	2
19	1	163688448	163	10
20	1	132509696	132	2
21	1	132509696	132	8
22	1	132509696	132	4
23	1	160145408	160	4
24	1	156602368	156	3
25	1	156602368	156	0

Table 9.1: Pixel counts and resolutions of the test imagery, along with the number of targets inserted into each image. Note that for the purposes of the experiment, all 2 metre resolution imagery was treated as if it was 1 metre resolution, albeit with a different range of clutter values.

[a] Assuming 1 metre resolution.

the same as a standard ROC[2]. In this case, the abscissa represents the average false alarm rate per unit area (km^2); *i.e.* for a given area, it represents the average false alarm rate. To convert to false alarms per unit area at each clutter level, the value on the abscissa should be multiplied by the number of targets that are classified at that clutter level.

The targets were randomly inserted at a rate chosen to match reasonable operational conditions of 1 target per 40 km^2 of imagery. Thus, about $108 \times 40 = 4,320$ km^2 of imagery was required for the experiment. Table 9.1 shows the number of targets that were contained in each of the 25 test images, along with their area.

Care was taken in the process of randomly placing the targets into the imagery. The human eye is particularly sensitive to edge effects. As a consequence, if the targets are not realistically inserted so as to avoid edge effects, then this will bias the experimental results. Furthermore,

[2] Strictly, the fraction of targets detected with sufficiently accurate localisation, as defined in the Localisation ROC.

all the target placements were checked to ensure that they did not occur in totally unreasonable situations, *e.g.*, on a cliff face. These details are more fully explained in subsection 9.2.9.

9.2.4 Preparation for the analysis-of-variance

As indicated in section 9.2.1, the experiment was originally designed to be a 3×3 factorial fixed effects design resulting in nine treatments. Upon this basis targets were inserted into the SAR imagery. However, the actual design used in the ANOVA was a 3×4 factorial fixed effects design, resulting in twelve treatments.

The reasons behind this are fully discussed in sections 9.2.7 to 9.2.9, later in this chapter, but a brief explanation will now be given. The twelve treatments were a combination of three levels of clutter and four levels of contrast[3] times target area product. It became necessary to introduce a fourth level of contrast (post experiment) because the experimental observers found targets in the imagery which could not be perceptually differentiated from the inserted targets, some of which were in a higher contrast range to that of the ranges set (low, medium, high) for the inserted targets. As the observers were not given any extra information which allowed them to differentiate these "extra" targets from the inserted ones, they were classified as real targets. Other non-inserted targets were discovered which were classified appropriately, as described in section 9.2.7.

As a result of the issues mentioned in the last paragraph, the data became non-ideal for the ANOVA approach, though they were accommodated in the ROC analysis by the FROC approach (section 9.2.3). The problem was that the amount of data at each combination of factors (treatments) was not equally distributed. This may cause problems in that each treatment condition is not guaranteed to contribute equally to the analysis, and this situation is likely to exacerbate the degree to which the assumptions underlying analysis-of-variance are violated (Keppel, 1991b).

There are approaches for dealing with this situation, such as the method of unweighted means or the method of weighted means, but these methods are flawed (Keppel, 1991c). The best approach, if sufficient data are available, is to randomly select the same number of samples for each treatment as exists for the smallest treatment sample. This approach was adopted in analysing the data shown in table 9.2. As shown in table 9.3, this resulted in a data set of greatly reduced size. However, an analysis-of-variance was carried out for both sets of data, with the analysis based on the second set acting as a check on the first analysis.

9.2.5 Experimental Procedure

A workstation, with a photometrically calibrated screen (*i.e.* that had its photometric output to grey-level input recorded) running the in-house analyst display software `displaytool` was used for the study. Changes were made to the software to log response times in conjunction with the observer's confidence rating for each detection. The observers were required to find targets

[3]Contrast and contrast ratio, are used synonymously in this chapter

	Contrast Level				
Clutter Level	low	medium	high	highest	row sum
low	455	174	10	10	649
medium	601	708	79	51	1439
high	174	640	199	201	1214
col sum	1230	1522	288	262	3302

Table 9.2: The number of targets at each combination of factors (treatment) for the experiment.

	Contrast Level				
Clutter Level	low	medium	high	highest	row sum
low	10	10	10	10	40
medium	10	10	10	10	40
high	10	10	10	10	40
col sum	30	30	30	30	120

Table 9.3: The number of targets at each treatment for the confirming ANOVA.

which had been pseudo-randomly placed into imagery according to the principles discussed in section 9.2.3.

The observers were instructed to search the swath by stepping through the image a screen at a time using non-overlapping image screens. During target logging, the observer placed the cursor over the selected target and hit a button causing the target coordinates to be recorded. Observers were then prompted to enter their confidence rating according to the 5-point scale outlined earlier (with an implied sixth rating of zero). No information on expected incidence or location of targets was given.

Each session was arranged for the same time every day for each particular observer and was limited to a maximum duration of one half-hour. Each observer was required to scan through one entire image per session in the order given in table 9.1. Thus, each observer typically sat through 25 sessions after an initial training session. (Some observers were quick enough to scan through multiple images in a single session.) The instructions given to the analysts can be found in Appendix L.

Some difficulties were experienced with observers interpreting instructions differently from my intention (see also section 9.2.7). While this emphasises the need to write instructions as clearly and unambiguously as possible, a component of the problem was the preconceived notions some observers had from previous SAR image analysis in an operational environment.

A full discussion of the experimental setup and its running can be found in sections 9.2.9–9.2.11.

9.2.6 Analyst Experience

Another issue is the experience level of the observers. If determining the effectiveness of expert analysts is the experimental aim, then only experts should be used in the experiments. However, ten experts were not available. These requirements could have been reduced with a corresponding increase in the amount of imagery each must search. However, this would have made the experiment even more onerous for each observer to carry out and would have reduced the willingness of observers to participate given their other commitments. Alternatively, if we wished to determine the effect of training on effectiveness, then observers with mixed experience would have been required, but more viewing would have been necessary to obtain statistical significance for the extra variable in the experiment. For example, if there were two levels of expertise, then the cost of the experiment would have been doubled over that indicated above.

Therefore, because of the increased cost that this question introduces into the experiment, I have not dealt definitively with it. The study used observers with mixed levels of experience and the results are dealt with as an average measure. As a result, these experiments indicate the performance level of analysts with skills that are somewhat below the level that could be expected with trained image analysts. In support of this pragmatic choice, the belief has been expressed by some that the detection task could be satisfactorily performed operationally by people who are not expert image analysts and who perform this task on a part-time basis only, along with their other responsibilities. However, this question cannot be answered with any confidence using the results of this experiment because the number of observers in each category would not produce statistically significant results. A further experiment designed to specifically answer this question would need to be run that fixed a number of the parameters that were varied in this experiment. However, this experiment addresses more fundamental questions regarding the "average" ability of observers to detect target-like objects in SAR imagery at varying contrast and clutter levels. In general, expertise is expected to have a multiplicative effect on the performance measured in this experiment.

9.2.7 Ground Truthing

Often, in visual detection task experiments, synthetic imagery is used that is well-defined and has some of the statistical character of real imagery, but none of its complexities. It differs from real imagery, which has cultural features that are distinct from the normal background texture and contains a variety of natural textures (synthetic imagery usually contains a limited variety of textures). This approach has not been possible in this experiment because of the need to make the experiment as close in character to an operational situation as is possible. Therefore I have used real imagery in my experiment.

Ideally, as mentioned in the introduction, I would have at my disposal imagery that was exhaustively ground-truthed. That is, every object in the scene of the image has been identified and logged, so that the cause of every target-like configuration of pixels in the image is known. If such imagery were available, it would have been possible in this experiment to test the ability of not just the analyst but of the whole SAR system to detect targets of military interest and

distinguish them from cultural clutter. I did not have such ground-truthing, and it was not feasible to collect it for the large amount of imagery used in this experiment.

The next best alternative to exhaustive ground-truthed imagery would be to scan through the imagery looking for every incidence of pre-existing target-like objects and then to send someone out into the field to determine whether or not the analysts are justified in calling it a target, or whether it should be classified as a miss or false alarm. Of course, any target on the ground that does not show up in the imagery would be missed. So with this approach, the performance of the SAR sensor to detect objects of interest would not be tested. However, the system's ability to distinguish target-like objects in the image from genuine targets of interest could still be determined by the experiment because we would know the cause of these pixel configurations. This would introduce a bias into the analysis in that the canonical list of pre-existing targets and false alarms would be prepared only from those found by the observers. It is conceivable that there would be pre-existing targets in the imagery that none of the observers detected, and, as a result, the reported probability of detection would be higher than it should be objectively. However, with a sufficient number of observers, it is a reasonable approach.

The approach just outlined of checking in the field was not feasible because of the cost and elapsed time since the imagery was collected. Therefore, the next compromise that had to be made was to consider visually ground-truthing the imagery. That is, visually compare all target-like objects detected by the observers (that was not artificially inserted there) and assess their similarity to the genuine targets that were inserted. If they were visually indistinguishable from genuine targets, then they were added to the list of inserted targets and not considered as false alarms. In reality, of course, some cultural clutter could fall within this category, *e.g.*, a fence post, but without ground truth it is impossible to take any other approach. This problem was minimised by selecting imagery with as little in the way of cultural features as possible, and this is consistent with the operational scenario anyway. Any small region of high cultural content, such as towns in the imagery, were marked out for observers to ignore, both to reduce the effective cultural clutter content and because the SAR role is to monitor remote areas.

In summary, it is important to understand the distinction between what should ideally be measured – the ability of analysts to detect targets of interest in a region *using* SAR imagery, and what was actually measured in this experiment because of the unavoidable compromises just discussed. This experiment measures the ability of analysts to detect target-like configurations of pixels in SAR imagery, and some of these objects will be due to genuine targets of interest and some will not be. As a consequence, it is not clear how to relate the false alarm rate reported here to that of the SAR system. Probability of detection is not so problematic, because experiments have been performed in the field to determine how known targets appear in imagery, and it is easy to measure how well you see an object in the imagery with a known location on the ground.

The assessment of the similarity of pre-existing objects found in the imagery by the observers to genuine targets should ideally be performed as a second experiment with a different set of observers. Again, time constraints and resource limitations have not allowed this, so a compromise was made by using a single expert (a colleague of the author), who divided all the detections made by the observers into 4 categories. These categories were: target of interest, a

permanent man-made installation (*e.g.*, building or power pole), a natural object (*e.g.*, a tree), or unknown.

The sensitivity of this vetting process to the categorisation was tested by assigning the unknowns to two different categories and running the analysis twice. The unknowns were those detections made by the observers that the expert could not clearly determine visually were either targets, installations or natural objects. Firstly, the unknowns were assigned to the target class, so in effect achieving the best-case false alarm rate. Secondly, the analysis was performed with the unknowns assigned to the class of natural objects (so that their detection is a false alarm).

Originally, it was planned to measure the classification component of the SAR analysis task, as outlined in section 9.1.1. However, the observers' responses to the questions in Appendix K indicate that very few observers did precisely what was wanted in this regard. This is due to two factors. Firstly, experienced observers tended to interpret what was wanted in terms of what their experience indicated. Secondly, because of the large number of instructions that were necessarily given to observers during initial training some details were forgotten. As a consequence, in the analysis of the results, anything that was identified as an installation was deleted from consideration, *i.e.* it was treated as neither a target nor a false alarm.

9.2.8 Compromises

This experiment measures the baseline performance for analysts using a SAR system in a search-mode of operation. It is only a baseline because of a number of compromises that had to be made to keep this experiment to a manageable size. The performance in an operational environment maybe expected to be better for a number of reasons, as discussed below.

Firstly, the 10 observers that were used had widely ranging levels of skill in the exploitation of SAR imagery for target detection. The level of skill varied from that of complete novice up to one observer who had been a professional image analyst at an earlier time in his career. As a consequence, the skill level of the observers as a whole was on average around the amateur image analyst level, so it maybe expected that suitable training would boost their performance.

Secondly, familiarity with a region would improve an analyst's ability to find targets.However, in this experiment, the learning factor was kept to a minimum by never using the same image twice. This was to ensure that the learning factor would not confound the results or cause an increase in the number of treatments investigated in the experiment. In this manner, this experiment purely measures target detection. It is expected that the learning factor will be investigated in a separate experiment.

Thirdly, analysts would normally have at their disposal more tools to examine the imagery than was allowed in this experiment. In particular, observers were not allowed to zoom in on a region of an image to examine it in closer detail. This would normally be used by analysts to increase their likelihood of making a correct decision by allowing them to examine carefully the edges of potential targets. Observers were denied the use of these tools for the simple reason that each level of option available to them is a multiplier on the size of the experiment, *e.g.*, if

there are 3 levels of zoom, then 30 observers would be needed instead of 10, or each of the 10 observers would have needed to sit through 75 images instead of 25. Despite these limitations, I believe that these issues are better examined in a small controlled experiment, designed to examine the improvements that result from each tool in isolation. These results then can be compared with the results of the base-line study described in this chapter.

Fourthly, analysts would normally have digital map information for the geographic region contained in the imagery.This was not provided to the observers in this experiment for the same reasons as above. This type of information would have the effect of enhancing an analyst's familiarity with a region and allow them to better distinguish existing infrastructure from targets. Previous imagery of a region would also normally be available to analysts to assist them in this process, this was not made available to them here.

Finally, the targets that were inserted in the imagery for this experiment were not inserted at realistic contrast levels. This was because it was necessary to test the full range of performance of the human visual system in the target detection context, covering every conceivable situation. It is necessary to interpret these results in the light of the operational target-background contrasts which have yet to be completely determined in further, operationally based, studies.

9.2.9 Preparing the Imagery

Ideally, in order to achieve a sufficient characterisation of an analyst's performance, a set of imagery with a range of clutters, containing known targets with a range of contrasts, is needed, but for the large amount of imagery required for this experiment, that was impractical. The approach taken was to choose swaths of existing imagery which were apparently largely free from targets (as per section 9.2.7) and artificially insert targets at randomly determined locations. This also allowed precise control of contrast, thus ensuring that a range of targets was present, varying from indistinguishable to obvious.

9.2.9.1 Extracting targets

A realistic set of candidate targets was obtained by extraction from existing images. A suitable data set was collected during a previous field trial. Two 4WD vehicles were positioned on a square of tarmac at the end of a runway at the Edinburgh RAAF base. The openness of the situation, with poor radar return from the tarmac, provided good target contrast. It was therefore simpler and more reliable to extract targets from this imagery than from images with targets in a more typical environment. The radar imaged the vehicles in spotlight mode on many runs from different directions. This provided a large number of images of suitable targets.

These swaths were processed with a program designed for processing strip SAR images. This does not properly account for the geometry of a spot SAR image. However, the distortion is not great and the appearance of targets is unlikely to be affected.

There were several stages in the process of producing a set of targets from the spot SAR images. Firstly, forty sections of the imagery, corresponding to separate images of the tarmac

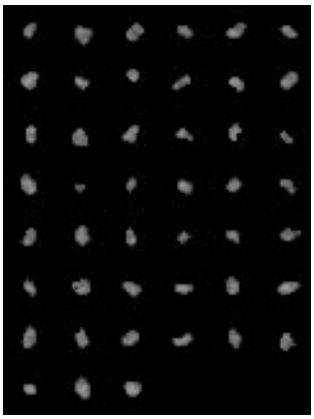


Figure 9.1: The set of targets used for insertion into the background images. The brightness is scaled to provide the desired contrast during target insertion.

region, were manually extracted. Although this introduces an undesirable manual process, which cannot be trivially reproduced, it would have been too much effort to do this stage automatically.

The position of the targets on the tarmac area was found by simple thresholding. A threshold was chosen that found many of the targets with no false positives. A small region around each target was then thresholded and cleaned with a simple neighbourhood filter.

The result of this process was forty five masked images of targets (see figure 9.1).

9.2.9.2 Choosing background imagery

Ideally, 1 m resolution background imagery was required. It was desirable that the imagery be of typical northern Australian regions containing few existing targets or other man-made features.

Unfortunately, neither of these requirements could be met. Most of the data available was at 2 m resolution (as wider swaths can be acquired at lower resolution), and there was insufficient 1 m data available. Also, while many swaths included remote regions, all had some man-made features present.

The data finally chosen were a mix of 1 and 2 m resolution imagery. Note that it is the image resolution that is being referred to here. The raw SAR data were collected with 1.5 or 3 m resolution complex samples. This was re-sampled to 1 or 2 m resolution when the image was formed.

The presence of existing man-made objects was accounted for in two ways. Firstly, all the data were viewed using `displaytool`, a locally written user interface program. Obvious, isolated targets and extended regions containing many radar brights [4] (such as towns) were logged. Secondly, a post screening was done on the analysts' detections (see section 9.2.11.1).

It would have been an onerous requirement for the analysts to log all the brights in a town. It would also have been inappropriate because the SAR system's requirement is to monitor

[4]The word "bright" is a jargon term for a group of pixels which stand out against the image background.

remote areas. Therefore, towns and other extended regions containing brights were marked on the imagery with a cross-hatch pattern. The analysts were specifically instructed to ignore these areas. Because `displaytool` has no facility for logging rectangular areas, an unusual method was used to log towns. The top left and bottom right hand corners were each logged twice. A program was then written that included a state machine to distinguish points from areas.

The `displaytool` program usually shows two versions of an image: one at full resolution and a "summary" image that is small enough to fit the entire region of the image on the screen. The summary image is usually constructed by dividing the image into blocks and averaging each block to produce a single pixel. To assist in finding radar brights, summary images were constructed, using the maximum of each block. With this change, `displaytool`'s threshold operation directly highlights radar brights of any size.

Several other options for dealing with existing radar brights were considered and rejected. Smoothing them out or pasting over them would be difficult without leaving obvious artifacts. Using very small sections of swaths or constraining the analyst's movement through the swath would make the analysts' procedure too artificial.

The images chosen were divided into approximately equal-sized chunks. The size of each chunk was chosen so than an analyst could reasonably be expected to complete finding targets in it in a half-hour.

9.2.9.3 Classification

To ensure that the targets were inserted evenly across different levels of background clutter, it was necessary to classify the background imagery according to clutter. As stated earlier, the measure selected for clutter is due to Waldman *et al.* (1988). The measure provides a clutter value for an image with a given step size in a given direction[5]. For the problem of distinguishing a target from background it is appropriate to choose a step size of the order of the target size. I used the average of the measures for the directions $(0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4})$ at a "standard" target width of 5 pixels.

As stated by Waldman *et al.*, a measure of clutter needs to be more specific than a measure of texture of the image around the target – the amount of texture at the scale of the targets of interest is the critical factor. The standard target dimension is less than the texture scale typically observed in SAR images. As a consequence, the Waldman *et al.* clutter measure reduces to a normalised weighting of the elements of the co-occurrence matrix, where the matrix is computed at the standard target size in each of the four directions (Waldman et al., 1988).

An efficient implementation of the Waldman *et al.* clutter measure was devised that avoids explicitly building the co-occurrence matrix. This is useful only because a small block size (64×64) surrounding each detection was used to compute the clutter level (section 9.2.1). Consequently, because the co-occurrence matrix is larger than this (256×256), it is sparsely filled. The matrix elements are usually weighted and summed; weighting and summing into

[5]See section 2.6.2 on page 58 of Chapter 2 for details.

Clutter	Area (km^2)
low	1121
medium	1180
high	1173

Table 9.4: The amount of area in each of the clutter regimes.

a single accumulator, as image values were examined, avoided building and looping over the matrix. With this and other coding efficiencies, the time to measure all the images was reduced from 15 days to 3 hours.

The background imagery was divided into small squares and the clutter and average power measured for each. A histogram of the clutter values shows a distribution with a small peak at low clutter, corresponding to shadow and water regions on the imagery, and a wide peak, corresponding to the remaining features. The wide peak was divided into three equal areas to set boundaries for different clutter regimes. Shadow and water were excluded from consideration, using thresholds on both clutter and power. The power threshold was -20 dB and the clutter thresholds were 0.121, 0.175 and 0.211. The amount of area (assuming one metre pixels) classified at each of the clutter levels was approximately equal and is shown in table 9.4.

Not unexpectedly, the clutter metric value varied with the image resolution. It was assumed that differences in visibility would be adequately reflected in the clutter metric values, so no further account was taken of resolution differences.

The screen luminance values (section 9.2.11.2) were not available when the targets were inserted. Instead, the approximate value of the contrast was found. The actual contrast was determined later and used in the analysis as discussed in section 9.2.11.3. The approximate contrast estimate corresponded well with visually perceived contrast in different background brightnesses and clutter levels.

A source of confusion is that there are four different "brightness" scales: the radiometrically corrected radar power, the contrast stretch, the screen pixel grey scale and the screen luminance. Here the screen pixel grey scale, which is linearly related to power, was used. Later (section 9.2.11.3), the actual luminance was used.

9.2.9.4 Target insertion

An interactive program was written to place the targets in the background imagery. For each required placement, a target was chosen at random and an image location was chosen at random from those of the correct clutter level. An image, showing the chosen location in context, was presented to the experimenter for confirmation and the details logged to a file. This required about ten minutes of interaction by the experimenter to place 108 targets.

Subsequently the actual insertion was done. The target pixels were scaled to the correct power to produce the desired contrast and the pixels inserted into the background imagery. The edge of the target was smoothed with the background to avoid an obvious step in brightness.

9.2.10 Running the Experiment

9.2.10.1 Displaytool

The program being developed as a model for the main user interface for the image analyst is called `displaytool`. Its main purpose is to allow the analyst to interactively view portions of an image, pan and zoom, mark manual detections and view and vet automatic detections.

Several enhancements were necessary for the program to be used in this experiment. The most important of these was a logging facility that recorded what part of the image the analyst was looking at and recorded the manual detections. Response times were also recorded, although they have not yet been used in the analysis.

Some further minor changes were made to enhance the usability of the tool for this task and to disable functions (such as contrast stretching and zooming) that were inappropriate.

9.2.10.2 Bookkeeping

To simplify the setting up required by each analyst during the running of the experiment, a point-and-click interface (figure 9.2) was provided. The analysts had simply to choose their name from the list and the tool determined which image they were up to and invoked `displaytool` appropriately. Details, such as ensuring detections were logged separately for each analyst, were taken care of automatically.

Figure 9.2: Front end program to prepare for analysts.

9.2.10.3 Instructions for analysts

The analysts received instructions on how to conduct the experiment by the following means.

- Written instructions detailing the use of `displaytool`, the desired method of scanning through the images and rating detections (see Appendix L).
- Verbal instructions with the same information and an opportunity for clarifying any unclear points.
- A demonstration and trial period with a specially prepared small image that included targets inserted at various contrast levels at known locations.

9.2.11 Preparing the Data for Analysis

9.2.11.1 Vetting

As discussed in section 9.2.7 it was necessary to visually ground-truth the imagery by vetting all the detections made by the observers. The approach taken was to view each detection (see figure 9.3) made by the analysts and manually classify it as a target of interest (*e.g.*, a 4WD), a permanent man made installation (*e.g.*, a building or pole), a natural object (*e.g.*, a tree) or unknown. Some obviously spurious detections were also deleted during this process (*e.g.*, one analyst placed a number of spurious detections in the corner of an image; presumably he wanted to see if we were awake).

Firstly, it was necessary to compare all of the ten analysts' detections and produce a list with all repeated detections elided. Of the 7360 detections, there were 3291 unique detections that were each classified in this way.

9.2.11.2 Measuring screen luminance

So that the analysts results' were meaningfully comparable, some care was taken that they were seeing the same thing. All of the analysts used the same machine; the monitor controls were set and then disabled to prevent changes during the experiment. The windows of the room were papered over and the lights left on to mitigate variations in lighting. The feature in `displaytool` that allows the mapping from image pixels to screen colour to be altered was disabled.

The screen luminance for each possible grey level was measured using a photometer. This ensures our results are meaningfully comparable with others in the literature. To achieve this, a full screen for each grey level was displayed and a luminance measurement taken with the photometer focussed at the centre of the screen. The results of these screen measurements, which were used in equation 9.1 to calculate the target contrast, can be found in table 9.5.

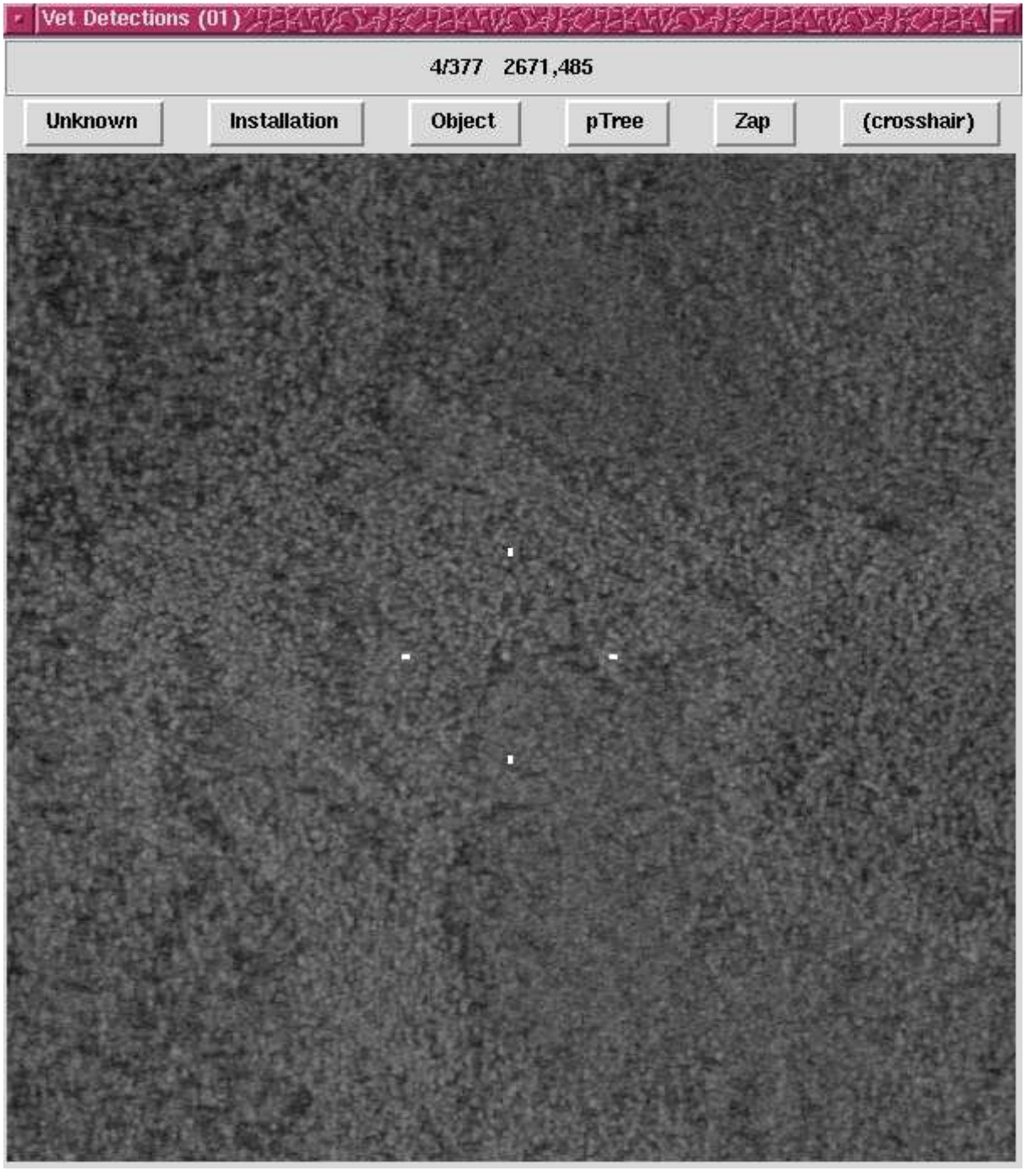


Figure 9.3: A screen shot of the vetting program. The user can categorise the target in the centre of the abbreviated cross-hair cursor, according to the buttons above the image.

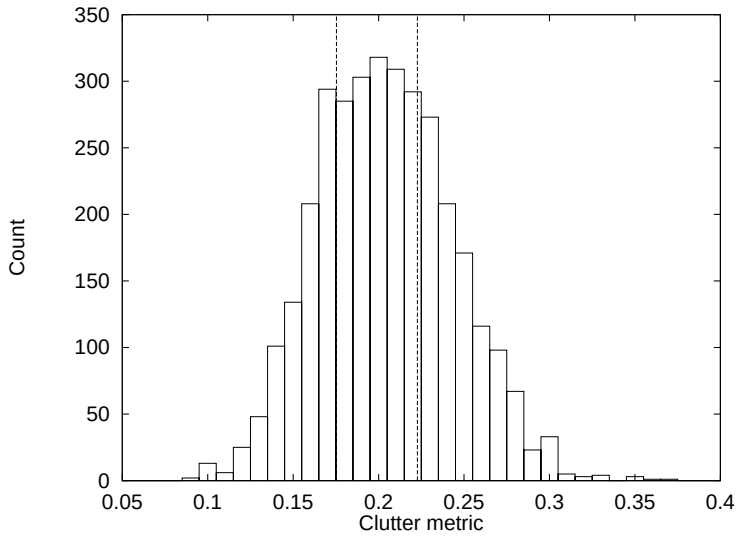


Figure 9.4: A histogram of the value of the Waldman *et al.* clutter metric centered on each of the detections made by the observers.
The division of the clutter values into the low, medium and high categories, according to equal areas under the histogram, is indicated by the dashed lines.

9.2.11.3 Measuring Target Contrast and Clutter

To construct the receiver operator curves the parameters required for each detection are: a measure of the contrast of the supposed target, a measure of the clutter of the region around the target and a classification of the detection as target or false alarm.

When the inserted targets were placed in the imagery the extent of the targets was known. The contrast, which depends on the extent of the target, could therefore be measured without difficulty. However, for measuring contrast at the detections, which in most cases, did not correspond with an inserted target, the extent of the supposed target was not known. Additionally, the measures used during insertion included the pixels "underneath" the targets.

These problems resulted in poor correspondence between the post-experiment measure and the pre-experiment classification for the inserted targets. However, all that was required for useful insertion of targets was for a sufficient spread of targets across different clutter and contrast regimes. Therefore, the numbers used during target insertion were abandoned and a classification based on measurements at detections were used.

A simple method was found to be sufficient for finding the target pixels. A 10×10 area, centred at the detection, was cut out. The brightest point in the area was found and the area was thresholded relative to the maximum grey-level (with a threshold of 0.9). Then those pixels remaining, which were contiguous with the brightest point, were deemed to constitute the target.

A histogram of the clutter surrounding each of the detections made by the observers is depicted in figure 9.4. Figure 9.5 displays a histogram of the target contrast-area product ca_{t} for each of the observers' detections.

What is required for the ROC is a determination of whether each detection is a target

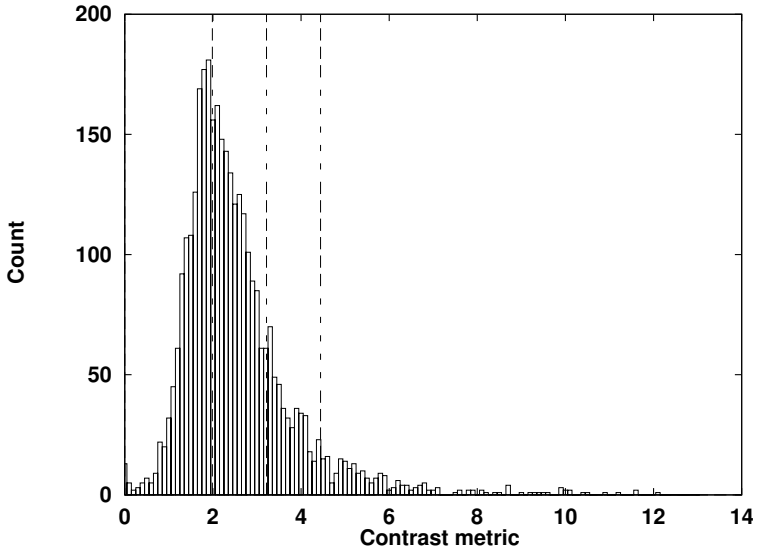


Figure 9.5: A histogram of the value of the target contrast-area product ca_t for each of the detections made by the observers. The division of the contrast-area product values into the low, medium, high and highest categories is indicated by the dashed lines. All the inserted targets fell into the lower three categories. A fourth category was necessary to account for the pre-existing objects in the imagery detected by the observers that were much brighter than the range used for the inserted targets.

or false alarm. For instance, it might be reasonable to treat permanent man-made objects as either targets or false alarms or to remove them from the analysis altogether. To allow the ROC analysis to treat these classifications flexibly, a filter was written to provide arbitrary mapping from the vetting classifications to "target", "false alarm" or "removed". In addition, individual images could be removed and the mapping could be controlled separately for different analysts to compensate for individual idiosyncrasies. See section 9.2.7 for the treatment used.

The analyst's detections for images 01 and 02 were discarded from the analysis because the observers were still rapidly changing in their performance, so these images were deemed to be part of training.

9.3 Results

The times recorded for each observer to search through all of the test images are given in table 9.6. This data possibly gives some comparative information on observer performance and differences in target distributions between images, but this information is too vague to add anything of value to the study. Therefore table 9.6 is not discussed further.

9.3.1 ROC Analysis

In ROC analysis the area under the curve, $A(z)$, provides a summary of the inherent discrimination performance obtained for the system of interest. This area can be interpreted as the average

value of detection probability on the corresponding ROC if the system's false alarm probability is selected randomly to be between zero and one. Equivalently, it can be considered as the average value of false alarm probability on the corresponding ROC if the system's detection probability is selected randomly to be between zero and one.

The results obtained for the ROC study are summarised in tables 9.7 and 9.8 which show the areas under the ROC's in figure 9.6. The ROC graphs have all been scaled to the same abscissa to facilitate visual comparison.

Four sets of ROC data were produced. Firstly two sets of analyst's response data were generated under different false alarm assumptions. In the first instance, any unknown targets that were detected by the analysts were recorded as a hit (best false alarm rate), while in the latter instance, any unknown targets detected were recorded as false alarms (worst false alarm rate). Then the sets of data were analysed in two ways to produce ROCs, and associated area under each curve, by using both a parametric (Gaussian) approach, (figure 9.6) and a non-parametric (with no form assumed for underlying distributions) approach (figure 9.7). As mentioned in section 9.2.3, in our case the abscissa represents the average false alarm rate per unit area (km^2); *i.e.* for a given area it represents the average false alarm rate. To convert to false alarms per unit area at each clutter level, the abscissa may be multiplied by the actual number of false targets that exist at that clutter level. The number of false targets in each clutter-contrast regime is depicted in table 9.9.

These analyses were performed on all four contrast and three clutter regimes. The four categories arise from the division of the range of contrast levels of the detections. Three levels were necessary to cover the treatment levels of the inserted targets, and an additional level was used to account for the pre-existing objects in the imagery that were deemed to be targets.

The areas obtained for both the parametric and non-parametric methods were similar, with a slightly higher estimation for $A(z)$ in the case of the former. This is usually the case since the non-parametric method fits straight line segments to the ROC graph while the parametric methods fits a smooth curve.

As expected, in all tables, the values of $A(z)$ increased with increase in target contrast. Also evident is a decrease in $A(z)$ with increase in clutter level, as determined by the Waldman *et al.* clutter metric. This was very apparent for the higher contrast levels. For the lowest range of contrast used, the $A(z)$ values obtained indicate performance only at chance; *i.e.* no discrimination was shown. This is because the lower target contrast level was deliberately chosen to be just beyond the perceptual range so as to ensure that the range was covered.

An apparent exception to this trend in $A(z)$ with clutter level is shown in the parametric tables (table 9.8), in the cells for medium and high clutter with contrast level at high. Here, the trend seems to be reversed, but it should be noticed that values are within the standard errors, rendering this apparent difference as unlikely. Moreover, perusal of their corresponding plots shows that the curves for medium clutter cross the curves for high clutter. These are examples of *improper ROC* curves (Egan, 1975a). In the standard ROC, the gradient of the curve at a

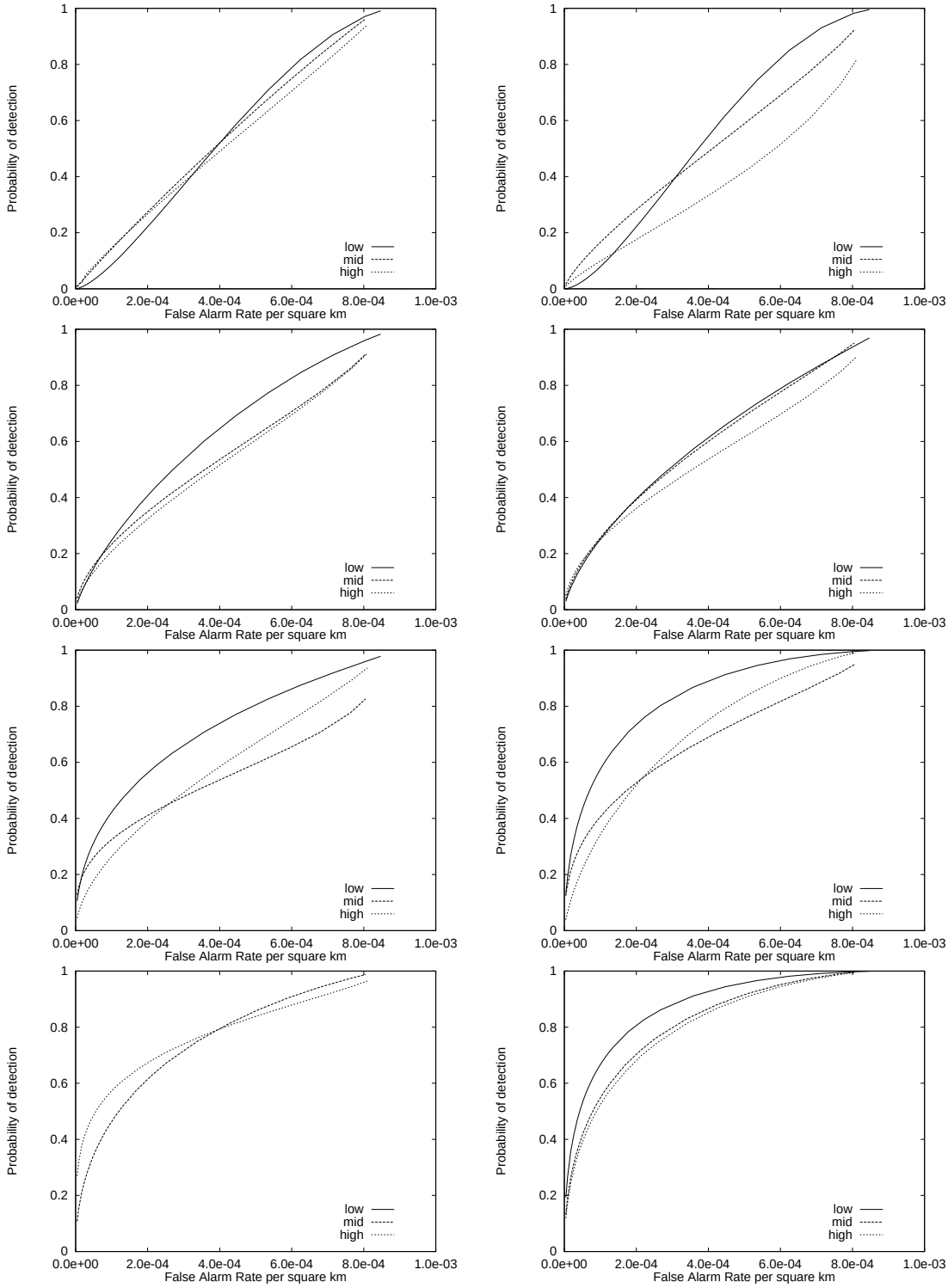


Figure 9.6: Parametric ROC curves. From top to bottom the plots are for low, medium, high and highest contrast. The left column is for best case false alarm rate (FAR); the right column is for worst case FAR. The legend labels: "low","mid" and "high" refer to low medium and high clutter levels respectively.

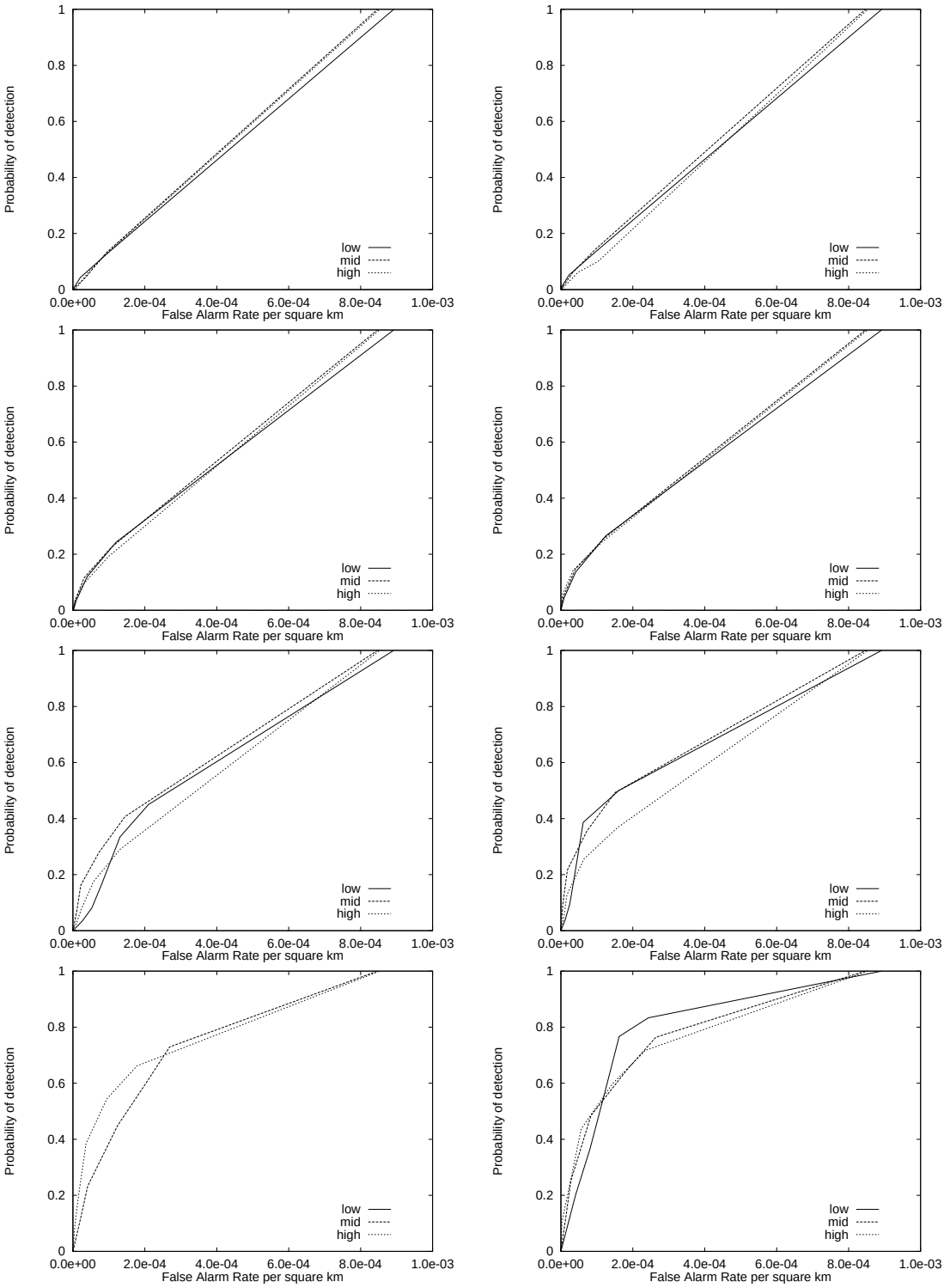


Figure 9.7: Nonparametric ROC curves. From top to bottom the plots are for low, medium, high and highest contrast. The left column is for best case false alarm rate (FAR); the right column is for worst case FAR.

point is equal to the likelihood ratio,

$$l(e) = \frac{f(e|s)}{f(e|n)}, \tag{9.2}$$

where $f(e|s)$ is the probability density of the evidence (e) for target present (signal), while $f(e|n)$ is the probability density of the evidence for target absent (noise)[6]. Therefore, the gradient of the ROC curve must be monotonically decreasing; *i.e.* no part of the ROC must ever be concave upward. However, this behaviour is evident in the curves just mentioned as well as the curves for the low contrast level. For further discussion see section 9.4.

9.3.2 Analysis of Variance

The aim of the experimental design used in the analysis-of-variance was to gain extra insights into the purely perceptual aspects of target detection in clutter. To this end, the approach outlined in section 9.2.4 was carried out, with no differentiation made between the detection of true or false targets.

The performance measure used for the analysis-of-variance was the probability of detection, or hit-rate, which is defined in equation (9.3). During the experimental sessions the observer's search time was also recorded as a potential performance measure. However, since the hit-rate varied from above 80% down to approximately 10%, I considered the use of search time as a measure of target detectability as invalid. The reasons for this are discussed in detail elsewhere (Woodruff and Newsam, 1994). However, in a nutshell, the search time is valid only when the variation in the hit-rate is not large.

The following probability of detection measure (hit-rate) was used:

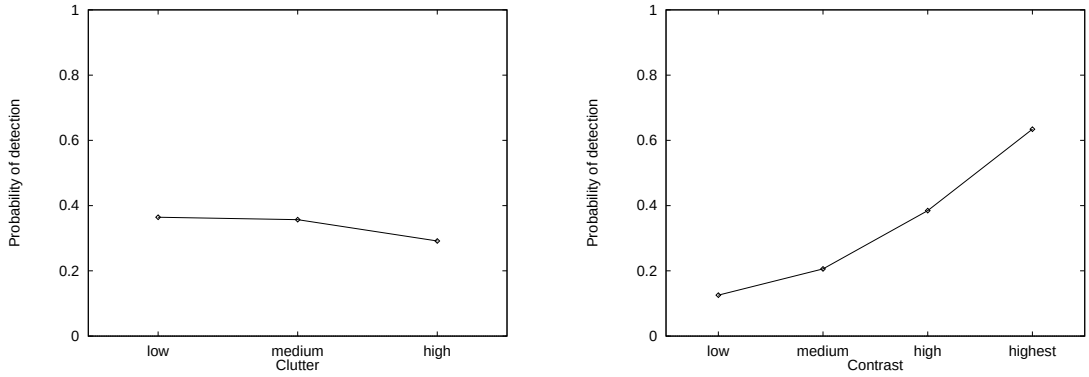
$$p_{di} = \frac{1}{N} \sum_{j} W_{ij} \tag{9.3}$$

where W_{ij} is zero for a miss, and one for a hit, for treatment i and subject j, and N is the number of subjects.

These data are shown graphically in figure 9.8 and figure 9.9. The former figure depicts the main effects in the analysis. As can be seen in figure 9.8(b), an increase in target contrast results in a concomitant increase in hit-rate with a statistical significance greater than the 99.999% level; *i.e.* less than a 0.0001% probability that this is purely a chance effect. Furthermore, an increase in the value of the clutter metric resulted in a concomitant reduction in hit-rate with a statistical significance at the 99.999% level.

The main effects only clearly indicate the relationship between the dependent and the independent variables in the absence of interactions. In this case they still provide summary information, but can hide more subtle effects. The next consideration is of the extent to which clutter and contrast levels interact. If no interaction exists, then the plot of hit-rate versus clutter level for each level of contrast, or vice-versa, will produce a family of parallel graphs.

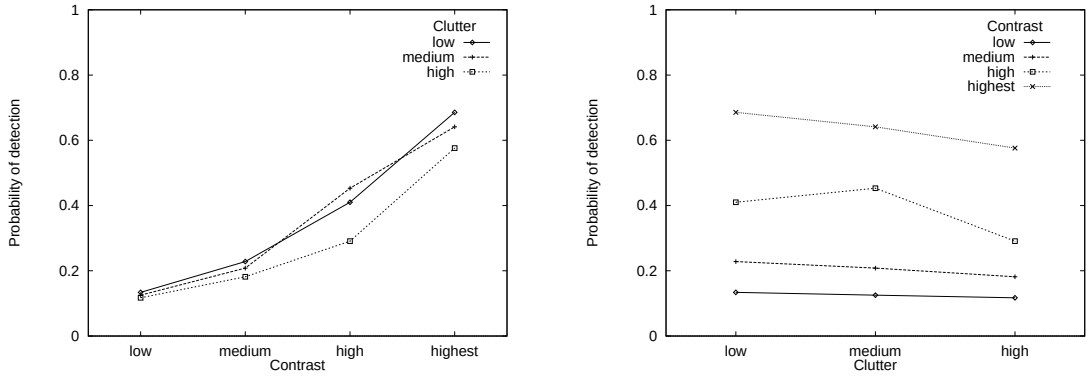
[6] See figure 3.2 on page 73 in Chapter 3, which represents the assumed underlying probability distributions.



(a) Effect of Clutter, $p < 0.00001$.

(b) Effect of Contrast, $p < 0.00001$.

Figure 9.8: The effects that the independent variables have directly on the hit rate.



(a) Hit-rate vs contrast, $p < 0.00001$.

(b) Hit-rate vs clutter, $p < 0.00001$.

Figure 9.9: The interaction between the independent variables.

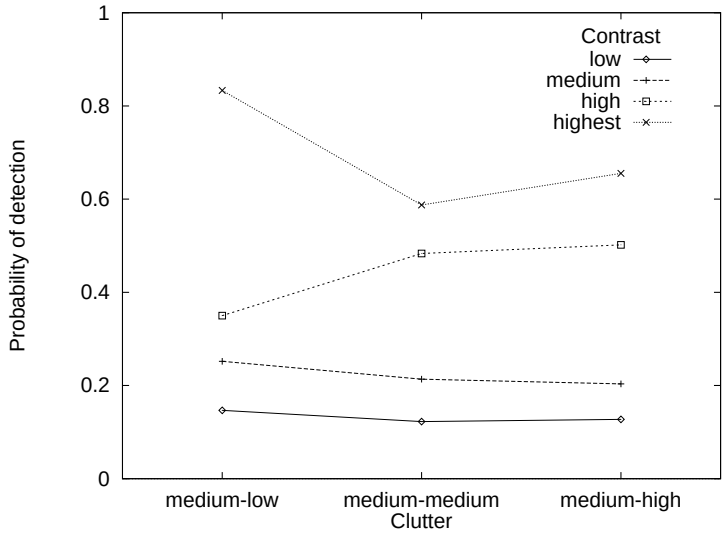


Figure 9.10: The interaction between the independent variables at finer scale.

These plots are shown in figure 9.9. As can be seen, the plots are not all parallel, which indicates an interaction occurred between clutter and contrast levels. This is confirmed by the ANOVA table which gives a significance level greater than 99.999% for the interaction between these factors; *i.e.* there is virtually no chance this is not a real effect.

Scrutiny of figure 9.9(b) reveals that there is little or no interaction at the low and medium levels of contrast, but there is obvious interaction at higher levels of contrast. At the highest level of contrast, the gradient of the approximately linear curve is steeper, indicating a stronger effect of clutter level on target detection, at this level. Interestingly, for a high contrast level, the curve departs from the linearly decreasing pattern that is exhibited for the other contrast levels. Here the hit-rate value for the medium clutter level is higher than that for the low level of clutter, going against the trend.

This interaction of clutter at the medium level with contrast was explored further by applying a finer scale to this clutter level, producing the levels: *medium-low, medium-medium* and *medium-high*. Figure 9.10 shows a plot of the interaction of medium clutter with contrast. The pattern shown earlier continues, with hit-rate increasing from medium-low to medium-medium at high contrast level, though the range has been narrowed to clutter metric values of 0.155–0.181. The plot for highest contrast and the three medium clutter ranges also shows a decrease in hit-rate at medium-medium clutter.

As indicated in section 9.2.4, these data were not ideal for analysis-of-variance, and a further analysis was performed on an optimised subset of the data. The results of this are summarised in table 9.11, as indicated earlier in this section. This table confirms the results obtained for the whole data set, showing both the contrast main effect and the clutter-contrast interaction as highly significant, even with this much smaller subset of data.

9.3.3 ANOVA Tables

A summary of the results for the analysis-of-variance is shown in tables 9.10 and 9.11, which are in the usual summary form of ANOVA results. Table 9.10 represents the results, using all the data that was also used for the ROC analysis, while table 9.11 presents the results obtained by the smaller subset of data optimised for ANOVA and designed to be used as a check on the results of analysis of the full set.

9.4 Discussion

It was demonstrated that that the degree of background clutter and target contrast had a (statistically significant) measurable effect on the detection performance of human observers. The problem is to define metrics to describe these factors quantitatively in a way that correlates with human performance. In the case of contrast, the Weber contrast metric (9.1) is usually employed. This however, is defined for simple plain targets, contrasted against plain backgrounds, and over a large but limited luminance range. In the absence of any obviously better alternative, however, the ca_t product was employed as described earlier in section 9.2.1, with c being the

Index	Power (dB)	Luminance (cd/m^2)	Index	Power (dB)	Luminance (cd/m^2)
0	-30.00	2.96	130	2.50	15.06
5	-28.75	3.07	134	3.50	16.01
9	-27.75	3.14	138	4.50	17.07
13	-26.75	3.19	142	5.50	18.26
17	-25.75	3.20	146	6.50	19.49
21	-24.75	3.21	150	7.50	20.74
25	-23.75	3.24	154	8.50	21.93
29	-22.75	3.27	158	9.50	23.25
33	-21.75	3.33	162	10.50	24.83
37	-20.75	3.41	166	11.50	26.18
41	-19.75	3.45	170	12.50	28.08
45	-18.75	3.71	175	13.75	29.52
49	-17.75	3.79	179	14.75	31.46
53	-16.75	3.94	183	15.75	33.16
57	-15.75	4.33	187	16.75	34.76
61	-14.75	4.48	191	17.75	36.68
65	-13.75	4.65	195	18.75	38.52
69	-12.75	5.04	199	19.75	40.44
73	-11.75	5.47	203	20.75	42.49
77	-10.75	5.85	207	21.75	44.43
81	-9.75	6.29	211	22.75	46.63
85	-8.75	6.77	215	23.75	48.47
90	-7.50	7.32	219	24.75	50.93
94	-6.50	7.83	223	25.75	53.01
98	-5.50	8.48	227	26.75	55.41
102	-4.50	9.12	231	27.75	57.62
106	-3.50	9.82	235	28.75	59.98
110	-2.50	10.64	239	29.75	62.48
114	-1.50	11.42	243	30.75	65.20
118	-0.50	12.25	247	31.75	67.46
122	0.50	13.09	251	32.75	70.38
126	1.50	13.98	255	33.75	73.85

Table 9.5: The grey levels, radiometrically corrected received radar power, and measured screen luminance of the experimental setup. The measurements were recorded with normal operating conditions with standard office lighting.

	Observer										
Image	1	2	3	4	5	6	7	8	9	10	Total
01	19	40	27	18	31	44	34	50	32	49	5:43
02	18	35	26	12	22	41	28	23	17	25	4:08
03	16	38	37	19	25	50	34	24	21	30	4:54
04	21	39	21	17	21	55	24	18	14	16	4:07
05	19	34	23	12	17	68	23	16	12	16	3:59
06	17	32	33	19	19	45	26	19	18	20	4:08
07	26	29	18	13	16	49	21	19	13	15	3:39
08	13	30	21	11	16	49	21	14	18	11	3:23
09	12	34	21	14	17	43	29	12	16	14	3:33
10	16	25	14	10	15	29	20	15	15	12	2:50
11	16	29	18	10	14	41	20	19	18	11	3:16
12	16	20	18	12	17	37	16	18	17	25	3:16
13	14	21	34	11	16	30	21	16	13	17	3:12
14	14	23	13	10	16	37	16	17	13	12	2:51
15	18	19	13	8	12	29	16	15	15	18	2:43
16	17	21	17	8	13	35	16	17	13	12	2:48
17	19	17	25	15	20	32	13	13	15	19	3:07
18	22	28	20	10	17	35	13	22	14	14	3:14
19	13	30	17	11	17	47	15	17	13	21	3:21
20	10	20	17	9	16	31	6	17	11	17	2:33
21	14	21	16	8	15	16	13	12	17	20	2:32
22	16	19	14	7	11	22	11	16	12	15	2:24
23	18	20	16	12	15	35	14	16	17	16	2:59
24	19	21	13	10	15	26	13	19	10	17	2:43
25	10	19	17	9	15	28	11	18	10	14	2:30
Total	6:55	11:04	8:28	4:57	7:05	15:53	7:52	7:44	6:25	7:35	83:55

Table 9.6: The rightmost column gives the sum of the times taken by all the observers to search through each particular image. The bottom row gives the sum of the times taken by each observer to search through all the images. The time is given in in hours and minutes. Note, the times in the main body have been rounded to the nearest second, while the total row and column each give sums of times to the nearest second,then rounded.

Clutter level	Contrast-Size product			
	low	medium	high	highest
low	0.5128 ± 0.0098	0.5578 ± 0.0139	$0.6716 \pm 0.0427^*$	–
med	0.5117 ± 0.0097	0.5547 ± 0.0082	$0.6352 \pm 0.0185^*$	0.7424 ± 0.0396
high	0.5135 ± 0.0170	0.5410 ± 0.0099	$0.5784 \pm 0.0142^*$	0.7694 ± 0.0162

(a) Best FAR

Clutter level	Contrast-Size product			
	low	medium	high	highest
low	0.5156 ± 0.0099	0.5673 ± 0.0129	$0.6931 \pm 0.0317^*$	$0.8700 \pm 0.5130^*$
med	0.5117 ± 0.0097	0.5629 ± 0.0079	$0.6751 \pm 0.0162^*$	$0.7781 \pm 0.0310^*$
high	0.5090 ± 0.0191	0.5602 ± 0.0099	$0.6032 \pm 0.0123^*$	0.7610 ± 0.0157

(b) Worst FAR

Table 9.7: Areas under the ROC curves (non-parametric analysis). A statistically significant difference (at the 95% level) in area for different clutter ranges with the same contrast range is denoted by an asterisk. The significance in difference in area was determined by the method of Hanley & Mcneil, 1982, which is equivalent to the Wilcoxon test.

Clutter level	Contrast-Size product			
	low	medium	high	highest
low	0.5533 ± 0.0351	0.6408 ± 0.0908	0.7218 ± 0.1076	-
med	0.5345 ± 0.0369	0.5453 ± 0.0641	0.5460 ± 0.0726	0.7525 ± 0.0565
high	0.5135 ± 0.1262	0.5313 ± 0.0857	0.5858 ± 0.0785	0.7730 ± 0.0471

(a) Best FAR

Clutter level	Contrast-Size product			
	low	medium	high	highest
low	0.5684 ± 0.0331	0.6194 ± 0.0748	0.8362 ± 0.1172	0.8750 ± 0.0953
med	0.5096 ± 0.0377	0.5957 ± 0.0415	0.6750 ± 0.0488	0.8102 ± 0.0342
high	0.4935 ± 0.0737	0.5313 ± 0.0857	0.7060 ± 0.0360	0.7730 ± 0.0471

(b) Worst FAR

Table 9.8: Areas under the ROC curves (Gaussian analysis).

	Contrast-Size product			
Clutter level	low	medium	high	highest
low	1400	540	30	-
medium	1670	1650	230	40
high	510	1100	430	160

(a) Best FAR

	Contrast-Size product			
Clutter level	low	medium	high	highest
low	2230	820	70	20
medium	2440	2940	470	60
high	850	2580	1290	300

(b) Worst FAR

Table 9.9: Numbers of false targets in each of the clutter and contrast regimes.

Summary of all Effects for ANOVA				
Effect	Degrees of Freedom (df)	Mean Square (MS)	F Ratio	p-level
clutter	2	0.7329	22.9924	0.00001
contrast	3	5.8903	184.7874	0.00001
interaction	6	0.1864	5.8471	0.00001
error	3290	0.0319	–	–

Table 9.10: ANOVA table for all the data.

Summary of all Effects for ANOVA				
Effect	Degrees of Freedom (df)	Mean Square (MS)	F Ratio	p-level
clutter	2	0.4533	5.6117	0.0048
contrast	3	1.7781	22.0146	0.00001
interaction	6	0.1799	2.2276	0.0458
error	108	0.0808	–	–

Table 9.11: ANOVA table for subset of the data with equal number of samples in each cell.

Weber contrast. Clutter, on the other hand, is a more complicated concept, and depends on target characteristics; *i.e.* what is clutter to one target is not to another. There exist many clutter metrics in the literature. However, as indicated in section 9.2.9.3, I chose the clutter metric of Waldman *et al.*, which showed promise during pilot studies with SAR imagery. The already large and complex experiment did not allow further clutter metrics to be explored in detail.

The results from the ROC and the ANOVA analyses showed that the contrast and clutter metrics used did indeed agree well with human performance as measured by the probability of detection. The metrics used have proven capable of describing the imagery in a quantitative way that allows the discrimination of certain salient image properties that apparently are also used by the human visual system.

It was learnt from the ROC analysis that the subjects were capable of a high degree of target discrimination under some circumstances. However, as expected, their performance, was very dependent on the target contrast and background clutter conditions, with the contrast being the dominant factor. As discussed earlier, the problem is that we do not have ground truth information, or enough information on the distribution of real target contrast levels, to relate our study to an operational setting. The levels of contrast required for good analyst performance were shown in this study to be quite high. However, this means little with respect to predictions of analyst performance in the operational setting in the absence of a typical contrast distribution in this context. This does, nevertheless, give some insight into human visual performance on target detection in clutter.

In regard to human visual performance, I discovered an interesting and unexpected phenomenon, as discussed in section 9.3. In general, analyst performance decreased with increase in clutter level. However, at the high contrast range, performance improved with targets embedded in high clutter backgrounds compared with targets in low background regions. This effect was shown to be highly significant in a statistical sense, but for this to be a real effect the clutter metric used must have behaved analogously to the human visual system. From the analyst's point of view, the apparent clutter at the point where performance improved must have been lower than that measured by the clutter metric used. This phenomenon requires further analysis and is beyond the scope of this study. I believe that this phenomenon may be due to the discrepancy between the assumed target size of 5×5 pixels and the actual size, which does vary around this ideal.

As discussed in section 9.3, for the ROC to be proper, the analyst must behave as a maximum likelihood observer. With the effect described in the last paragraph, however, this criterion would not have been met in all situations. The observer is in effect setting the value of the likelihood function (9.2) based on the evidence available when determining whether a target exists in the image at a particular location. This evidence is really based on a subjective metric set up in the perceptual space in the observer's mind. However, it is also based on the physical evidence for a target in the image itself, which includes its clutter characteristics. Therefore, if the clutter metric was measuring a value for clutter that is inappropriate for the human observer, then in the context of the ROC analysis, this would appear as an irrational decision

rule on the part of the observer at this point. This would result in an improper ROC. As has already been observed, this appears to be the case. It is most evident in the parametric ROC curves at the high contrast level, as would be expected. The non-parametric analysis seems to be more robust to departures by the subject from the performance expected from a maximum likelihood observer.

Notwithstanding this possibility, it seems more likely that the improper ROC curves derived from the parametric analysis at the lower contrast levels occurred for other reasons. At these levels of contrast, the observer performance was only at about chance level. That is, the probability distributions underlying the ROC curves are best described as Bernoulli distributions (Egan, 1975b). However, the parametric method inappropriately fits a binormal distribution (via a maximum likelihood fit) to the data. This, I conjecture, is the cause of these poorly fitted curves. This situation was probably exacerbated by the unequal counts in each rating category for detections versus those for false alarms.

9.5 Conclusions

This study has shown how difficult it is to set up an experiment to reach a good compromise between internal and external validity considerations; *i.e.* to reach a usable tradeoff between laboratory control and real world applicability. Some problems here resulted in non-optimal data sets, which caused some ROC curve fitting problems, particularly for the parametric curve fitting algorithm. Both the non-parametric and parametric methods agreed well in all cases when proper ROCs were obtained. This indicates that in complex experimental regimes, as used here, the non-parametric approach is the more robust.

The experimental data indicate that SAR image analysts performed well only with relatively high contrast targets in the context of clutter, and at this level they obtained a 76% to 87% chance of a correct decision at highest typical target contrasts. The analysts' performance depended on the clutter level, as measured by the Waldman *et al.* clutter metric. This metric has been demonstrated to correlate well with human perception of clutter, as observed in SAR imagery, in a rigorous way. This also adds weight to the findings on the "localness" of clutter as determined in the experiment described in Chapter 7.

Due to the lack of appropriate data, the relationship between the contrast distribution of real targets of interest and that of our experimental regime has not been investigated here. In addition to the results presented here, this crucial piece of information is needed in order to relate this study to the performance of image analysts using a SAR system in an operational context.

Chapter 10

Conclusion and Summary

The aim of this thesis was to consolidate and comment on the image measure literatures, to find, through experiment, the salient properties of electronically displayed real world complex imagery that influence human performance in well specified visual tasks (of real relevance), and from the data collected consider the most effective application of image measures to this imagery for the prediction of human performance.

10.1 Summary of Results.

An introduction to certain aspects of image quality measures was provided. Image quality was defined in the context of this thesis and clutter metrics were related into this concept. A very brief and basic introduction to the human visual system was given with some basic models.

10.1.1 Image Measures

An analysis was given of image measures which were classified according to those features they were designed to quantify. The relevant literature was reviewed and image measures were categorised according to their underlying principles and the intended mode of application. They were also classified according to the spatial extent of application; *i.e.* as local or global measures, and some measures, such as edge measures, were found to be intrinsically local in nature. These measures were also classified as either similarity (fidelity) or interpretability (intelligibility) measures. Clutter measures were regarded as a form of the latter. It was noted that these classification schemes were mutually inclusive.

The analysis of the image quality measure literature found a natural classification of image metrics according to the image features they are attempting to quantify. Five basic classes of image measures emerged. They are listed here and were discussed in detail in Chapter 2.

(i) L^p norm type measures, of which the most common is the mean square error $[L^2]$ (MSE);

(ii) Modulation transfer function (MTF) type measures. These are commonly used in assessing imaging systems, but can be used in assessing images directly;

(iii) Information measures;

(iv) Decision theoretic measures;

(v) Signal detection measures.

The above listed measures are usually applied in a global sense, but could be applied with local support; *i.e.* calculated over a localised area of the image. The definition of "local", in the application of image measures, was explored experimentally in Chapter 7, and is summarised later in this section. There exist another two classes of measures, which are usually applied to local features. These are "edge quality measures" and "texture measures", where texture measures are frequently in the class of the entropy based measures; *i.e.* they are often an information measure.

In the area of image analysis, there appeared to be two main avenues of attack on the application of measures to images.

(i) a statistical or feature based approach and

(ii) a syntactic (or structural) approach.

The literature on image quality measures is pervaded with the former type of measures, but very little if any has been done, using the latter approach, in this context. This is developed in the section 10.2 on further work in this chapter.

The term "clutter" was introduced and is used as a general term to describe spatial and, sometimes, spatiotemporal variations in imagery which reduce the availability of target information to a specific sensor. Although other researchers in the area of clutter are interested in man-made electro-optical sensors, this thesis was concerned with the HVS as the sensor.

The image quality measure and clutter measure literatures appear as largely distinct.However, this thesis argues that clutter measures are in fact not different to the image measures, except they include a meta metric to characterise target as distinct from background; *i.e.* the clutter metric has explicitly built into it the concept of distinguishing target from non-target (background). It was shown that clutter metrics are in fact a form of interpretability or intelligibility metric, as found in the image quality literature, except for the extra condition just discussed.

10.1.2 Image Similarity

A study was described which investigated how humans related to the same scene information presented both as infra-red (IR) imagery and optical imagery. As this was my first psychovisual experiment, I investigated the applicability of a subjective methodology for producing an interval scale as a metric of image similarity, and applied some basic image quality metrics to get a feel for their application.

From this experiment it was apparent that the content of the image scenes was important in how an image was perceived. It has been indicated that relationships of the regions within the

scene were very important, as was seen with the regions of shadow and sunlight. These relationships could not be measured by means of global statistical measures alone. It became apparent that, to capture the complexity of the images, measures of local (region-based) image properties are required and to "measure" the relationships between image objects, syntactic[1] (Gonzalez and Wintz, 1987a) types of measure were probably required. The latter type of measure is beyond the scope of this thesis. However, a proposed system for image quality assessment, which uses this type of measure, is discussed later in this chapter in the context of further work.

This study convinced me that further experiments should be based on well defined visual tasks, such as detection or recognition, in order to determine directly the effects of image properties on performance; *i.e.* to determine the image quality in terms of *utility* rather than in terms of nebulous qualities such as aesthetics.

10.1.3 Still Image Compression

With the greater integration of computers and telecommunications, and with their increasing demands on digital storage and transmission systems, image compression is becoming increasingly important. In many cases the overriding concern is the quality of the reconstructed image. Therefore there is a need to answer the question of how much compression can be achieved (what is the minimum image quality required) to achieve a certain task, in the context of constraints upon image storage and/or transmission. Consequently, when considering an imaging system for a well defined visual task, such as surveillance, it is important to first assess to what extent the compression module is likely to affect user performance on the given tasks.

To this end, my next psychophysical experiment was based on a very well controlled, though applicable, visual task using a variety of potential performance measures. The aims of this experiment included that of analysing human visual performance under a well defined visual task, with imagery that has undergone degradation. Image compression was chosen as the means of degradation because it can be precisely controlled and image compression has practical application.

Two methods of static image compression – JPEG and a fractal-based method – were compared in terms of the detectability of simple targets following compression and decompression of the images containing such targets. Targets consisted of rectangles of various sizes and contrasts, which were embedded in images of natural terrain. Using compression ratios of from zero to thirty five, it was found that the loss in detectability of targets in images compressed using the fractal technique was significantly greater than the loss for the JPEG-compressed images. In contrast to this finding, at the time of the experiment, fractal compression was generally considered to be superior to other methods in the achievement of higher compression ratios. It was found that, for compression ratios < 10, there is no significant difference in performance between the compression schemes, as overall visual performance at these compression levels is little degraded.

[1] That is structural and relational aspects of the objects within the image scene itself.

The results from a reliability analysis showed that the experimental technique used was highly reliable, particularly when group rather than individual data was used. The same basic techniques were then applied with confidence in later experiments. The "hit-rate", or probability of detection, was found to be more robust, though less sensitive, than response time as a measure of performance.

10.1.4 Video Compression

A study on the effects of video compression on visual tasks was performed. The work described in the last sub-section studied the effect of still image compression on target detection. This work continued in a similar vein, with an examination of the effects of video compression on target recognition. A set of video compression experiments was performed which required observers to recognise ships in randomly presented video sequences. The sequences had controlled levels of contrast and multiplicative noise, and were compressed and de-compressed at a variety of compression levels using MPEG-2 encoding under standard settings.

The data demonstrated a clear effect of video compression on observer performance in target recognition. However, this effect was not large enough to cause serious misclassification of targets and was largely due to the degradation caused at the maximum video compression (2.0 Mbps). The degradation caused by compression did increase the time required for target recognition, particularly at the highest compression level. In a real world setting, depending on the application, this may introduce an intolerable degradation in observer performance.

This experiment highlighted some aspects of individual differences in perception, at this slightly higher (than detection) task level. It became clear that similarity or otherwise of targets was very subjective. This information was mainly gleaned from the post experimental debriefing of the observers, but it was also evident in the data. Despite this, from both anecdotal and statistical evidence, all observers found one class of target the hardest to discriminate from each of the other targets. This suggested that the features of this target overlapped with the features of the other three classes of target in perceptual space. Further exploration of this would require techniques such as Multi-Dimensional Scaling (Evans and Attaya, 1978).

There was a significant learning effect discovered, even after the initial training sessions. In other words, the observer's performance kept improving with practice. This meant that the HVS was learning the distortions introduced by compression and either compensating for them or modifying the internal model of the target. This has serious implications for the training of personnel using systems which require compressed imagery, as it shows that operator efficiency can be greatly improved by training in performing surveillance type tasks.

The upper limit for MPEG-2 compression obtainable was at 2.0 Mbps (approximately 80:1) and this was the level accounting for most of the degradation. A new standard for low bit rate video compression called MPEG-4 (ISO/IEC, 1998) has just emerged. This standard allows much lower bit rates to be obtained, and has been designed for many applications, including surveillance systems. When software and/or hardware to enable MPEG-4 encoding is available, further similar experiments need to be performed to evaluate this new standard, in a visual

task mode, since unlike MPEG-1 and MPEG-2, it is most likely to be used in this way; *i.e.* for surveillance, remote medicine *etc.*

Some of the sequences produced were extremely degraded after compression and had added noise and contrast reduction (prior to compression), so that individual still frames were quite difficult, if not impossible, to interpret on their own. This suggested that the temporal processing of the human visual system is a powerful aid in detection and recognition, and is further motivation for pursuing this research.

Implications for Task Related Video Quality Metrics

This work has some general implications for video quality metrics, with particular reference to video compression. Firstly the temporal dimension has to be considered. Even though this aspect of the work was only indicative here, it was clear that temporal processing in the HVS made a significant difference in the ability of observers to classify targets. Therefore, any video quality or utility metric must be applied over several video frames, and be applied both spatially and temporally. The number of frames that need to be "sampled", and the relative weighting of spatial and temporal properties to be included in a metric, probably depends on the visual task and application, and are important subjects for further research.

In considering a metric which will predict human performance, cognisance has to be taken of the learning effect demonstrated here. Since vastly different performances can be achieved for the same video stimuli at different points in the learning curve, an appropriate metric must allow both for the type of visual task and the level of observer experience. This may mean that basic metrics work under a "meta" metric. (See section 10.2 in this chapter, which discusses such a system.)

This study has shown that, when dealing with higher level visual tasks such as target classification and recognition, as compared with simple detection, individual variability is greater. Though this is not unexpected, it was interesting here how each individual observer rated the similarity between the test targets. Depending on the particular "quality" being measured, the subjective variability in the perception of similarity may impact on the design and application of the quality metric. As an example, consider a metric which is required to produce a classifiability index of various video sequences. This metric would produce this index by measuring appropriate physical video image characteristics and applying a suitable mathematical construct to map the target's features into a space, analogous to the human perceptual space, in which the targets are separated. However, there appears to be some variability in how humans map the target features into perceptual space and these need to be considered in the metric design.

Possible Future Directions.

The type and amount of information required for various visual tasks itself varies. Therefore, it makes sense to consider optimising the compression scheme for specific tasks or classes of task (*i.e.* to produce the minimum information needed for a specified level of task performance). In

the case of JPEG and MPEG, this may be achieved by adapting the entropy coding and/or the quantisation tables in the standards; existing implementations usually come with default tables which have been optimised for viewing natural scenery.

Other authors (Lohscheller, 1984; Watson, 1993) have considered this aspect to some extent by optimising the JPEG quantisation matrix for specific images. They do this by varying separately each of the discrete cosine transform basis functions to determine when a subjective just-noticeable-difference is obtained. This seems to minimise the apparent compression artefacts. The main interest of those authors, however, was in aesthetic image quality, which often has little relation to image quality in a task-oriented sense.

One useful avenue for development would be to consider adjusting the JPEG quantisation tables in a systematic way in order to ascertain the effect on well-defined visual tasks such as target detection. A similar approach has been proposed in medical imaging (Kostas et al., 1993), but the quantisation adjustment was not well defined in this small study. If this approach turns out to be useful with JPEG it could also be applied to MPEG, including MPEG-4, but with some qualifications. Firstly, the extra temporal dimension and the use the human visual system makes of it, will come into play; this may for instance, require different quantisation tables from those used for still images. Secondly, there are more free parameters to be considered and optimised in the MPEG protocols, such as the effects of varying the motion compensation regime.

10.1.5 Localisation of Clutter

In theory, properties of clutter can be defined globally or locally. However, in the literature, the distinction between local and global clutter is arbitrary. If the image contains different clutter types, global clutter metrics may be inappropriate and are expensive to compute. In the literature, the standard approach of setting the local domain of the clutter metric to twice the expected target size is adopted without any justification. However, it was found that the size of the local clutter region around a target has a strong effect on the probability of detection of that target, and that this was affected by regions much larger than twice the target size. It was also discovered that this effect was much stronger for targets subtending less than 0.8 degrees of visual angle than for larger targets. In the case of the former, the fall-off in human visual performance with clutter region size was approximately quadratic compared to a slight linear fall-off for larger targets.

A model was presented explaining these phenomena, indicating that the auto-covariance function characterising the clutter is the main determinant of the size of the region of local clutter, but is reduced for larger targets.

Further Work

The research reported here did not elucidate the detailed stimulus interactions across multiple receptive fields and further work needs to be done to clarify the mechanisms involved.

This study considered only a narrow class of clutter and simple target type. Alhough more work needs to be done, even for this situation, in order to fully understand the mechanisms involved, other clutter and target scenarios need to be considered in order to arrive at appropriate clutter measures for practical application.

10.1.6 A New Image Metric (GEM)

This work was aimed at producing an image metric capable of measuring the effects of image processing parameters on the clinical value of Single Photon Emission Computed Tomography (SPECT) images, which are used in the Nuclear Medicine departments of hospitals. The overall long-term aim was to develop an automatic system for optimal image filter parameter adjustment.

A measure, called the *gradient energy measure* (GEM), for quantifying the effect of filtering on SPECT images, was developed and evaluated. This proved to be a reliable measure of image smoothing and noise level, which, in preliminary studies, agreed with human perception. Both the Laplacian and Sobel versions worked well, though the former correlated slightly better with subjective data. The Sobel filter was more insensitive to noise, which may make it more suitable in some situations. The fact that the GEM correlated well with subjective ratings means that it could possibly be used as a stand-alone measure, though this is not necessary for its intended use in optimising SPECT image quality using full-blown subjective experiments.

Further Work

There is a model of HVS function that appears to be modelled to a first order by the GEM. This model is known as the "energy integrator" model (Green and Swets, 1966c; Moulden et al., 1990). More work needs to be done to determine if this is an adequate model, and this may be an interesting area for further research.

This measure was designed to be used in conjunction with subjective analysis of SPECT images in order to find the optimum filter parameters in terms of clinical image quality. This will require a large subjective study in order to to ascertain the optimum set of processing parameters and to calibrate the GEM. It is recommended that the receiver operating characteristic (ROC) method is the most appropriate, as trade-offs between hit-rate and false-alarm-rate are important in the clinical setting, though some insight, with less subjective data, could be obtained from using an analysis-of-variance (ANOVA).

10.1.7 Human Target Detection Performance in Clutter

The effect of clutter, in Synthetic Aperture Radar (SAR) derived images, on the performance of human image analysts in the surveillance context, was determined in terms of the analyst's receiver operating characteristic (ROC). The experiment was designed to correspond as closely as possible to the expected real world operational mode of analysts using similar imagery. In par-

ticular, the effects of target contrast and background clutter on human analyst target detection performance were quantified.

This study has shown how difficult it is to set up an experiment to reach a good compromise between internal and external validity considerations; *i.e.* to reach a usable tradeoff between laboratory control and real world applicability. Some problems here resulted in non-optimal data sets, which caused some ROC curve fitting problems, particularly for the parametric curve fitting algorithm. Both the non-parametric and parametric methods agreed well in all cases when proper ROCs were obtained. This indicates that, in the complex experimental regimes used here, the non-parametric approach is the more robust.

The experimental data indicated that SAR image analysts performed well only with relatively high contrast targets in the context of clutter, but at this level they performed quite well; *e.g.*, they performed with a 76% to 87% chance of a correct decision at highest typical target contrasts. The analysts' performance was dependent on the clutter level as measured by the Waldman *et al.* clutter metric. This metric has been demonstrated in a rigorous way to correlate well with human perception of clutter, as observed in SAR imagery. This was not applied in the usual way of setting its region of support to twice the expected target size. Based upon the work described in Chapter 7, the local area was defined to be about 3° around the target.

Further Work

Further work needs to be done to test various clutter metric types on imagery from different sensors. A controlled study should be done using these metrics with the conventional setup, *i.e.* with the local region set to twice the target size, and comparing the ability to predict human performance with the metrics that have the local region extended as determined in this thesis.

10.2 Longer Term Further Work

Further work that needs to be done in the context of the actual studies performed for this thesis has been articulated in the previous section. This section goes on to propose some additional work. Although based on the research as described in this thesis, this future work is more visionary in nature and will require considerable effort and time to fulfil.

The research that has been done in the area of image quality measurement is very multidisciplinary in nature. Not only has this work been published in the journals of diverse disciplines, but performing the research also requires knowledge across a broad range of disciplines and sub-disciplines; *e.g.*, pure and applied mathematics, statistics, physics, engineering, computer sciences (including AI), psychophysics, psychology and physiology. Therefore, the problem of image quality evaluation is very daunting indeed, not only because of its multidisciplinary nature, but because image quality is a ubiquitous concept in any domain requiring human interpretation of images, which makes a general definition of image quality an elusive goal.

To recapitulate the discussion in the introductory chapter, there was little effort in image quality research until the mid 1970's. This effort waned by the end of that decade, and remained at a low level until the mid eighties. There was then a short spurt of activity, followed by further decay to a low level. Due to developments in image compression and communications technologies, there has been renewed interest in image quality measures in recent years. However, there has been only limited success, and in many studies, the quality *figure-of-merit* used does not correlate well with the subjective evaluation of image quality. Two major reasons for this are:

(i) No allowance in the quality model for the response of the human visual system (or a lack of understanding of the HVS);

(ii) Ignorance of the purpose for which the human observer is using the image(s).

With respect to item (ii), it is known that subjective judgements of image quality depend on the purpose of the image. It may be expected, therefore, that the relative importance of the various physical image parameters in determining subjective assessment of image quality should depend on the task in which the image is used. Further, one may expect that the combination of physical image parameters that are optimal for an applied task, will depend on the task considered.

These matters make it questionable as to whether image quality can be expressed in a single general measure. Thus, most authors have attempted to define a measure of some aspect or aspects of image quality, quite often in the context of some specific tasks. However, there have been some attempts to quantify image quality in a single measure or metric, and these have met with limited success (as discussed in Part I of this thesis).

These attempts to develop image quality measures have been ad hoc, applying measures to address specific aspects. There has not been any attempt, as was the aim of this thesis, to develop a systematic approach to measuring image quality by applying a suite of measures appropriate to the situation. Notably absent in the image quality literature is any mention of syntactic based measures[2] rather than purely statistically based measures. Part of the motivation for using these measures comes from the experience gained from the experiments in Part II of this thesis.

10.2.1 A System for Image Quality Estimation

From these experiments, it is apparent that the content of the image scene is important in how the image is perceived. Relationships of the regions within the scene are very important. These relationships could not be measured by statistical means alone.

Therefore, a system is proposed, which includes the various types of information needed to predict the quality (utility) of an image for human use. Figure 10.1 shows the proposed functional block diagram of a system to perform image quality evaluations. This system is complex, and it would be impractical to attempt to develop all the system requirements in the

[2] measures concerned with the spatial relationship of objects within images.

short to medium term. The main problem area is the syntactic analysis, with the requirement for automatic image segmentation and object recognition, which is an entire research field in itself. A system, with a less ambitious design, and which requires a human to perform the image segmentation, is presented later. However, the fully automatic system would be suitable as a long term research goal for a team of researchers.

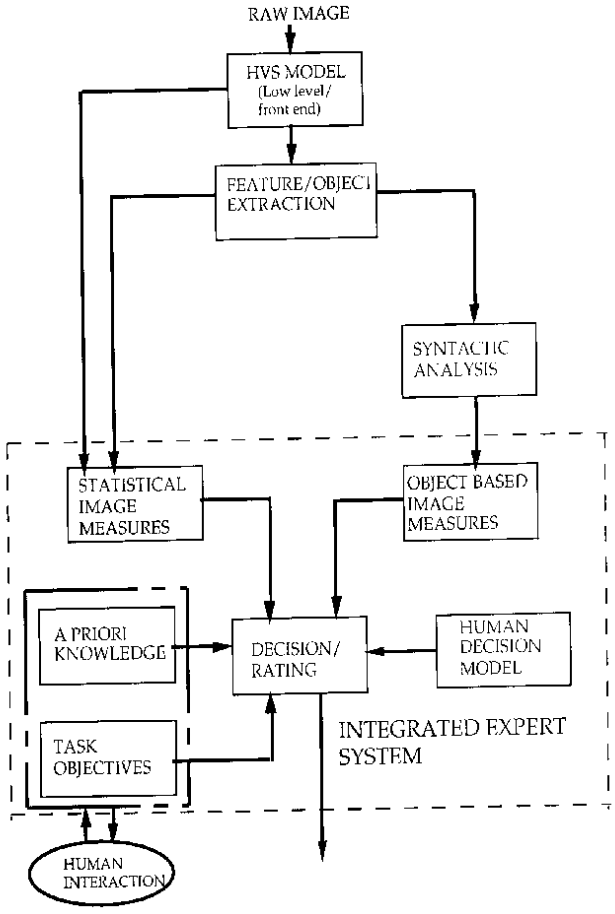


Figure 10.1: Ideal System for Image Quality Evaluation.

System Components

A basic description of the system components is as follows:

- **HVS model**: Simple HVS models were discussed in section 1.2.3 in Chapter 1. Here, the raw image data is operated on by the HVS model to emphasise and de-emphasise appropriately the raw image data to conform to human perception.
- **Object extraction**: Here, the image is broken up into its constituent *objects*. The objects themselves can be characterised by their *features*. These features can be statistical or structural in character. This is the most difficult task within the system and requires image segmentation, object detection, object recognition and feature selection and extraction. All

these areas have not been achieved successfully in practical systems and are under intense research worldwide.

- **Syntactic analysis**: In the *syntactic* approach to image analysis the image is decomposed into simpler sub-images. By recursively continuing the process, the image is finally broken up into a set of *primitives*. Objects in the image are represented by these primitives and the relationships between them. These primitives are represented by a set of symbols and their relationships described by a *syntax*, such as connectivity rules.
- **Object-based image measures**: These are measures based on the syntactic analysis of the image. The spatial and contextual relationships between the images are considered in order to define measures on the image which relate to human perception and evaluation of the image.
- **Statistical image measures**: These are measures which are evaluated at the pixel level. All image quality measures in the literature seem to be of this type. Some of these measures already have a model of the HVS built into them.
- **Human decision model**: In order to simulate the human decision process a model, based on psychovisual experiments is required, to process the low-level image input data.
- **A priori knowledge**: This includes contextual and other knowledge that an experienced observer would have about objects in scenes.
- **Task objectives**: This is another part of the knowledge base, but could include interactive input. Specific knowledge, about the type of objects and particular contexts, would be included here.

In this system the subset of measures used for a specific task would be automatically selected from the set of measures implemented.

10.2.2 A Realisable System

To reduce the complexity in the development of the system to a realistic level, the following modifications will be made (see figure 10.2):

- The object extraction will be performed interactively by a human operator;
- A relatively small set of image measures will be selected and then be fixed;

The human decision model will be implemented by a neural network (NN), and will probably be trained using back-propagation. The training input will consist of subjective ratings and/or measures of human task performance, such as response latency or results of Receiver Operating Characteristic (ROC) analysis.

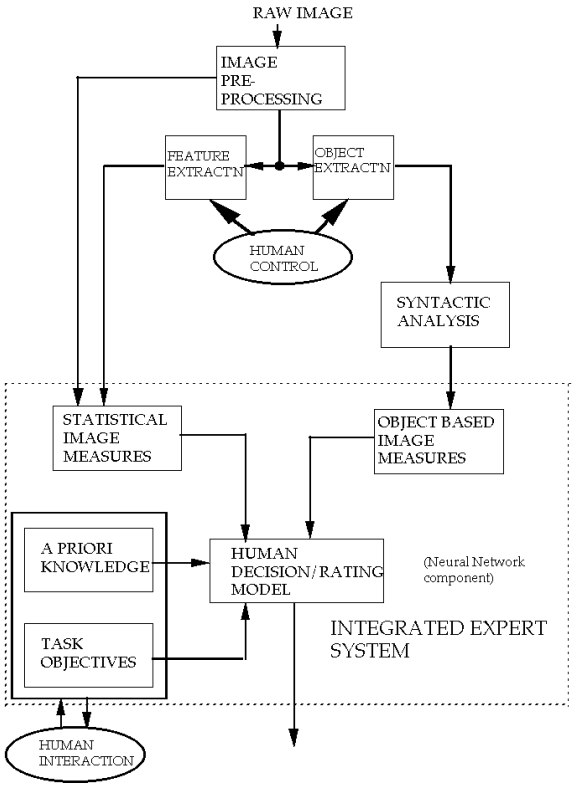


Figure 10.2: Realisable System for Image Quality Evaluation.

Anticipated Use of the System

An image, say a synthesised scene, will be input to the system. Following this, a set of global statistical measures will be performed on the image.

With the aid of (say) a mouse, the operator will then interactively segment the image, defining objects of interest semantically. Using this new data, the system will perform a syntactic analysis and compute measures, including statistical measures on selected objects and/or features. The system will then interact with the operator about the specific tasks for which the image will be used. An output in terms of a quantitative scale will be presented, along with a qualitative description of the reasons for the image quality rating on this scale.

Appendix A

Radiometric & Photometric Quantities.

Radiometry and photometry are concerned with describing the origination and transfer of electromagnetic (EM) energy from the scene to a sensor system. There is a one-to-one correspondence between radiometric and photometric quantities, but the latter evaluates the effectiveness of EM radiation in stimulating the human eye.

Figure A.1 shows a sphere of 1 metre (m) radius. The area (A) of the sphere $= 4\pi R^2$; at $R = 1\,m, \quad A = 4\pi\,m^2$. By definition, a sphere subtends a solid angle of 4π steradians (sr). Therefore, the area on the surface of the $1\,m$ sphere, covered by the intersection of the solid angle of 1 steradian $= \frac{4\pi}{4\pi}\,m^2 = 1\,m^2$.

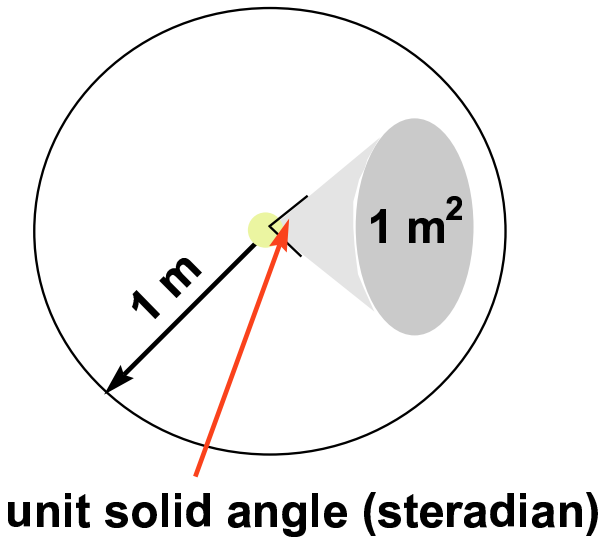


Figure A.1: The cone of unit solid angle, subtends an area of 1 m^2 at the surface of a 1 metre sphere.

Photometry

Photometry, as distinct from radiometry, allows for the spectral sensitivities of the human visual system; *i.e.* it considers light that is visible to the human eye. For light, a point source has an intensity of 1 candela (600,000 int candles), if at a distance of $1\,m$ from the source, the amount of flux through an area of $1\,m^2 = 1$ lumen. The flux (θ) through the solid angle (Ω) is $\theta = \int_\Omega I d\omega$. The luminous intensity, $I = \frac{d\phi}{d\omega}$, where I has of units watts/steradian (W/sr).

Luminance

Equivalent to spectral radiance in radiometric units. Units = candelas/m^2 or W/sr m^2.

Transmittance

$T = \frac{\phi_t}{\phi_I}$ where: ϕ_t = transmitted flux and ϕ_I = incident flux.

Reflectance

$R = \frac{\phi_R}{\phi_I}$ where: ϕ_R = reflected flux. Note: Only for Lambertian reflectors.

Optical density

$d_t = \log 1/T = -\log T$ (transmitted light)
$d_r = -\log R$ (reflected light)

Illuminance (E)

$$E = \frac{I \cos\theta}{R^2};$$

i.e. with 1 candela intensity source, the illumination at $1\,m$ over an area of $1\,m^2$ (by 1 lumen of flux) at normal incidence.

Radiometry

$(\forall\lambda)$,

$\Phi_e(\lambda)$ the *spectral radiant flux*, total flux

$$\Phi_e = \int \Phi_e(\lambda) d\lambda, \text{ watts.} \tag{A.1}$$

spectral Radiant Flux

$$I_e(\lambda) = \frac{d\phi_e(\lambda)}{d\omega_e} \tag{A.2}$$

Unit = W/sr. For an isotropic source $\phi_e(\lambda) = 4\pi I_e(\lambda)$ and for an extended source area dA, the emittance $M_e(\lambda) = \frac{d\phi_e(\lambda)}{dA}$. Allow for viewing angle and intensity of dA.

Spectral Radiance ($L_e(\lambda)$)

$$L_e(\lambda) = \frac{dI_e(\lambda)}{\cos\theta_e dA} = \frac{d^2\phi_e(\lambda)}{\cos\phi_e dA d\omega_e}. \tag{A.3}$$

If the emitter is Lambertian, then $L_e(\lambda) = 1/\pi M_e(\lambda)$ W/sr m^2.

Spectral Irradiance ($E_e(\lambda)$)

The flux falling on an external surface is given by $E_e(\lambda)$.

Solid angle = $\frac{\text{area}}{\text{R}^2}$

$$\begin{aligned} \Rightarrow d\omega &= \frac{\cos\theta dA}{R^2} \\ \Rightarrow R^2 &= \frac{\cos\theta_e dA_e}{d\omega_e} \\ &= \frac{\cos\theta_r dA_r}{d\omega_r} \end{aligned}$$

from A.2

$$\begin{aligned} R^2 &= \frac{\cos\theta_e dA_e I_e(\lambda)}{d\phi_e(\lambda)} \\ &= \frac{\cos\theta_r dA_r I_e(\lambda)}{d\phi_e(\lambda)} \\ &= \frac{\cos\theta_r I_e(\lambda)}{E_e(\lambda)} \end{aligned}$$

$$\Rightarrow E_e(\lambda) = \frac{I_e(\lambda)\cos\theta_r}{R^2}$$

Appendix B

Confusion Matrices

In order to determine any measure of classifiability, it is necessary to assess the accuracy of the classification procedure. This will allow a degree of confidence to be given to those results.

A *confusion matrix*, as shown in table B.1, indicates the number of pixels classified incorrectly; *i.e.* indicates to which class the pixels are classified and shows erroneous results. It is

		Reference classes							
		A_1	A_2	A_3	A_4	A_5	Total	% Correct	% Commission
	A_1	35	2	0	3	0	40	88	12
Test	A_2	1	42	6	0	3	52	81	19
image	A_3	2	0	16	0	6	24	67	33
	A_4	5	12	4	112	0	133	84	16
	A_5	0	0	2	3	96	101	95	5
	total	43	56	28	118	105	350		
	% 0mission	19	25	43	5	8			

Table B.1: A Confusion Matrix for Five Classes.

common to average the percentage of correct classifications as a measure of the overall classification accuracy van der Lubbe (1984a) In this case it is 83%. However, a more appropriate measure may be to weight the % correct classifications in each class according to areas represented on the ground-truth map (or some *a priori* probabilities).

Let the confusion matrix consist of elements a_{ij} which represent the number of pixels of class A_j that have been classified as class A_i. Three types of (mis)classification may be distinguished:-

(i) Correctly classified pixels, which are found along the diagonal of the confusion matrix and are denoted by a_{jj}

(ii) The pixels that have erroneously been excluded from the j^{th} class (omissions). For the reference class A_j this is

$$\sum_{i,i\neq j} a_i$$

(iii) The pixels that have been assigned to the wrong class (commissions). For m class A_i of the test image that is:-

$$\sum_{j,i\neq j} a_j$$

Clearly, in the case of perfect classification, the confusion matrix is non-zero only on the principal diagonal.

Appendix C

Pre-Experiment Statistical Analysis for ANOVA

In this appendix we present the steps that are necessary to determine the number of observations required to guarantee a specified level of significance and power of the statistical test used in the experimental design. In our experiment the means for the various treatments (determined by particular combinations of parameter values) are compared to decide whether they are significantly different statistically. For testing the equality of two means (μ_0 and μ_1), the null and alternative hypotheses are $H_0 : \mu_1 = \mu_0$ and $H_a : \mu_1 \neq \mu_0$, respectively, while the two types of decision errors that can be made in hypothesis testing are:

Type I error Rejection of the null hypothesis when it is in fact true. The probability of making this error is usually denoted by the symbol α.

Type II error Acceptance of the null hypothesis when it is in fact false. The probability of making this error when a specific alternative is true is usually denoted by the symbol β.

Making a type I error results in accepting the alternative hypothesis when it is false and thus making inflated claims, while making a type II error will lead to reporting no significant results when in fact they exist. A conservative approach is to make the probability, α, of a type I error smaller than the probability, β, of a type II error. The *power* of the experiment is defined as $1 - \beta$; *e.g.*, if β= 0.1, then power is 0.9, and this means that if there is an effect, it will be detected 90% of the time.

Effect Size An important consideration is the *effect size* that is required to be detected. The effect size is the degree of change in response produced by a change in the level of a factor. This can be specified in the same units as the factor, and is designated by δ; *e.g.*, in a sound level discrimination test, this could be measured in dB. It is often advantageous to have a dimensionless measure of effect size. In the case of testing the effects of two treatments this is usually defined by a parameter d, *viz.*:

$$d = \frac{|m_{\mathrm{a}} - m_{\mathrm{b}}|}{\sigma} \tag{C.1}$$

where m_a and m_b are the mean responses for the two factor levels, $\delta = |m_a - m_b|$, σ is the standard deviation of the responses (assumed equal for both levels) and is in the same units as m_a and m_b. Thus d can be thought of as a measure of the distance between or the amount of overlap in the two response distributions. This parameter is employed in t-test tables for comparing the means of two sample sets.

When more than two sample means (say k) are compared as in an ANOVA, d can no longer be used (except to give an upper estimation of sample size, see below), so another measure of effect size is defined;

$$f = \frac{\sigma_m}{\sigma} \tag{C.2}$$

where σ_m is the standard deviation of the response means and σ is the common standard deviation as before. This parameter is the effect size index tabulated in F-test tables. Further information is needed to specify f; since there are more than two means, their relative positions on the f line ($f \to [0, \infty)$) can be distributed in various patterns. This pattern must also be specified. Here it is assumed that the k means are distributed evenly over the range of the responses. Then f is defined (in terms of d) as:

$$f = \frac{d}{2}\sqrt{\frac{k+1}{3(k-1)}}. \tag{C.3}$$

Typical vales for the parameters are as follows: $\alpha = 0.05$, which means that the significance level is set at 95%; and $\beta = 0.1$, that is, the power is 90%; and finally the estimation error is set at $\delta = 0.25\sigma^2$.

To get an initial estimate of the number of observations required (N), we use the expression

$$N = 2(Z_\alpha + Z_\beta)^2 \frac{\alpha^2}{\beta^2} \tag{C.4}$$

where Z_α is the $200\alpha\%$ point of the standard normal distribution and is Z_β the $200\beta\%$ point of the standard normal distribution. The number of degrees of freedom (df) are then calculated, $\mathsf{df} = 2N - 2$, and N recalculated by using the Student's t-distribution.

The following table indicates the number of samples or observations required per treatment for different values of the sensitivity index and experimental power. These values are for comparing two means only and form an upper limit on the number of observations required per treatment when comparing more than two means; *i.e.* when the number of means is greater than two, the requirements for the number of samples is reduced. The significance is set at 95% level.

	Sensitivity index (d)			
Power	0.2	0.3	0.4	0.5
0.9	524	230	142	84
0.8	392	172	106	63
0.7	308	135	83	50
0.6	244	107	66	40

It is obvious that power and sensitivity are inter-dependent. We have designed the experiment to pick up the smallest possible effect within the limitations of our resources. The actual level of the effects in the experiments determines the power.

The sample size requirements obviously impact on the costs in setting up, running and analysing the experiment. The impact on analysis is relatively minor.

Appendix D

Target Parameters in Chapter 5

This appendix describes the procedure used to select values for target variables. The variables of interest are target area (=size) and target contrast. Target contrast was defined by:

$$c = \frac{\mu_T - \mu_B}{\mu_B} \tag{D.1}$$

where μ_T was the mean target luminance and μ_B was the mean local background luminance. Target area was manipulated through a variation of a linear scaling factor, the target size factor, s_t. The target was a rectangle with a fixed aspect ratio, so that when normalised ($s_t = 1$), the target was 9 pixels high by 4 pixels wide. These dimensions were scaled by a factor of $\sqrt{(}s_t)$; *i.e.* target area was equal to the nearest integer value of $36 \times s_t$ pixels.

Target size and target contrast were random variables, uniformly distributed within predetermined ranges (items (i) and (ii) respectively), which were subject to a constraint on the product of target contrast and target area, which is defined by the inequality in (D.2). A pilot experiment was designed to estimate the maximum and minimum values to most efficiently cover the perceptual space for the full compression experiment, and from this, ranges of values for each variable were obtained.

(i) Target size factor ($\propto$ area) $0.7 \leq s_t \leq 2.75$;

(ii) Target contrast $0.2 \leq c \leq 0.6$

The size-contrast constraint was:

$$0.4 \leq s_t \times c \leq 1.7. \tag{D.2}$$

Note that, having chosen a value for tc, luminance values were converted to greylevels.

Appendix E

Target Insertion Procedure

The main difficulty in inserting artificial targets into real imagery is in making the correct adjustments to pixels on sides of the edges of the target. These adjustments are needed to simulate the sensor blurring that would have occurred around the edges of a real target. When imaged by a system having a point spread function (PSF) $S(u,v)$[1], a background with radiance $B(x,y)$ will produce an image $I(x,y)$ where I is the convolution $S * B$. If a target with luminance $T(x,y)$, covering a region R (with zero luminance outside R), is inserted into the original background, then the system will produce the modified image

$$I' = S * (B - B_R + T). \tag{E.1}$$

Here B_R denotes the function obtained by windowing B to R. That is, B_R has radiance B in the region R and zero radiance outside R. The blur introduced by the PSF means that pixels in I' that are close to, but not in, R will be affected by the change in the background over R, and so differ from the corresponding pixels in I. Likewise, pixels in I' that are in R cannot be automatically set to have radiance values $T(x,y)$ due to the contribution by blurring from that part of the background radiance field not obscured by the target. If the original image I could be deconvolved to recover the true background B, then the new image I' could be calculated by (E.1). Deconvolution is well known to be a difficult problem, so this approach was not followed here. Instead, the following simpler approximate procedure was adopted. The detectors are assumed to be square, have a 100% fill factor, and have a sensitivity that is uniform throughout. It is also assumed that $S(u,v)$ is normalised to have unit volume, so that the total intensity of a blurred image is the same as the original radiance field. We replace each of the continuous quantities I, B, and S by the discrete quantities i, b, and s corresponding to the values of the continuous field at each pixel centre, so that we have to deal with digital images only.

Finally, given the region R (which may or may not have boundaries that coincide exactly with the edges of pixels), we define the windowed digital image b_R by setting $[b_R]_n = \alpha(n)[b]_n$, where $\alpha(n)$ is the proportion of the n^{th} pixel lying in R. Next consider the constant image e, in which every pixel has a value of one, and window it to give the uniform target. Let $f = s * e_R$ be the convolution of this digitised target with the digitised PSF; *i.e.* f is the image of a uniform

[1] See section 2.2.4 in Chapter 2 for a definition of PSF

target against a black background. The algorithm for inserting a uniform target with brightness μ into the image i now proceeds as follows:

(i) Calculate the mean grey level μ_R of the background within the target region R, and form the windowed image $i' = (i - \mu_R f) - (i - \mu_R f)_R$. This excises that part of the background that would be obscured by the target and attempts to remove, from pixels lying outside the target region, the contribution from the (approximately known) obscured background.

(ii) Calculate the convolution $s * i$ and window it to the region R, to give the approximate contribution $(s * i')_R$ to image pixels within the target from the background outside the target.

(iii) Finally calculate the image $i_\mu = i' + (s * i')_R + \mu f$.

This procedure has been applied here to generate images of rectangular targets of uniform radiance and linear dimensions

Finally the following procedure was used to set target contrast to a specific value C.

(i) Compute the mean grey level L_B in the background in the region corresponding to the target region R plus all pixels whose distance from R is less than half the diameter of the support of the PSF.

(ii) Determine the mean grey level L_R over the region R of the image i_μ where $\mu = L_B + 100$. The contrast in this image is:

$$C(\mu) = \frac{L_R - L_B}{L_B}, \text{ the Weber contrast.}$$

(iii) Now to produce an image with a given contrast C, set

$$\mu(C) = L_B + \delta L, \text{ where } \delta L = \frac{100C}{C(\mu)}.$$

Target insertion repeated with this value of $\mu(C)$ will give an image $I_{\mu(C)}$ in which the contrast between the target and its background will be C as required.

Appendix F

Technical Problems in Setting up the Experiment in Chapter 6

Setting up this experiment required considerable effort and some hard lessons were learnt. A major problem was the logistics of the experiment, in particular the requirement to compress, decompress, load onto tape and then display in the order of 5000 video sequences (512 for each of the 10 observers in a different random order). Initially, it was intended to perform the experiment by playing the sequences from a Panasonic AJ-D350 D3 format digital video cassette recorder. The initial analog video was converted to digital, then processed on computer with the plan to lay the sequences down on the D3 tape. The MPEG2 encoding was done using software and this was first thought likely to be the major bottleneck. However, migrating the software from a PC to various UNIX workstations brought the estimated processing time down to a reasonable 50 hours of CPU time (although disk space storage was a problem at times). The major problem turned out to be the laying down of the processed sequences onto D3 tape. This was done by existing special purpose software which loaded digitised sequences from a PC disk onto the D3 via its RS232 communications port. Due to mechanical constraints, this process turned out to take approximately 20 secs per frame independent of the number of frames to be down-loaded at any one time, making it impossible to load all 512 sequences in any reasonable time. (The method was sufficient, however, to set up a small pilot experiment to test experimental design.) Therefore another, previously unavailable, solution was sought in the form of one of the recently released MPEG2 PC boards, which allow direct playing of MPEG1 & 2 bit streams from hard disk. This not only removed the need to lay down sequences on tape, it also greatly reduced the storage problem and allowed much greater software control of the experiment. There were however, considerable difficulties in sorting out "teething" problems with the software for driving the MPEG2 board.

Appendix G

Derivation of the Fractal Image Simulation Algorithm for Chapter 7

Consider a fractal Gaussian random field (as defined on page 132 of Chapter 7) within a square domain h, defined by vertices at [(0,0),(0,1),(1,1),(1,0)], which is our pixel size. Since fractal GRF statistics are scale invariant, this does not lose any generality.

We are given that

$$C(r) = kr^{2\delta}, \tag{G.1}$$

while in general

$$C_h(\mu,\nu) = \left\langle \int_0^1 \int_0^1 L(x,y)dxdy \int_0^1 \int_0^1 L(x'+\mu, y'+\nu)dx'dy' \right\rangle, \tag{G.2}$$

where $r = \sqrt{\mu^2+\nu^2}$ and $< \cdot >$ is the expectation operator and $x' = x + \delta_x, \quad x' = y + \delta_y$.

Substitute (G.1) into (G.2).

$$\Rightarrow C_h(\mu,\nu) = \int_0^1 \int_0^1 \int_0^1 \int_0^1 k([x'+\mu-x]^2 + [y'+\nu-y]^2)^{\delta} dxdx'dydy', \tag{G.3}$$

$$\Rightarrow C_h(\mu,\nu) = k\int_0^1 \int_0^1 ([r_x-\mu]^2 + [r_y-\nu]^2(1-|r_x|)^{\delta}(1-|r_y|)dr_x dr_y, \tag{G.4}$$

since $\int_0^1 \int_0^1 f(x-x')dx'dx = \int_0^1 f(r_x)(1-|r_x|)dr_x$ (Chapple, 1997).

(G.5)

It was shown by Chappel [*ibid*], that (G.3) becomes

$$\begin{aligned} C_h(\mu,\nu) = {} & k\int_{-1}^1 [\ln|r_x-\mu+(r_x-\mu)^2+(r_y-\nu)^2|]_{r_x=-1}^1 dr_y \\ & -\frac{k/2}{1+\delta}\int_{-1}^1 ([r_x-\mu]^2+[r_y-\nu]^2)^{1+\delta}]_{r_x=0}^1(1-|r_y|)dr_y \\ & +\frac{k/2}{1+\delta}\int_{-1}^1 ([r_x-\mu]^2+[r_y-\nu]^2)^{1+\delta}]_{r_x=-1}^0(1-|r_y|)dr_y, \end{aligned} \tag{G.6}$$

and that the integrals can be evaluted from -1/2 to + 1/2 since

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x' - x) dx' dx = \int_{-1}^{1} f(u)(1 - |u|) du. \tag{G.7}$$

Appendix H

Instructions to Observers in Study Detailed in Chapter 7

Introduction

The proposed work deals with the effects of clutter on human target acquisition. If we consider an image of a scene containing a target, clutter is here defined as any structure(s) in the image apart from the target, which masks the target or confuses the observer as to the location and/or class of the target. Properties of clutter can be defined globally or locally. If the image contains different clutter types, global clutter metrics may be inappropriate and are expensive to compute. It is also likely that, in terms of detection, rather than search, local clutter is more important. The problem is to determine how local is "local"? In the literature, a standard approach of setting the local domain to twice the expected target size is adopted without any justification. This work addresses this issue.

Aims

The aims of this research are to achieve the following listed items by gaining knowledge of the extent and functional form of the effects of local clutter on human target detection.

- Give more accurate prediction of human target detection, since clutter metrics will be more representative of human visual responses;
- Increase the efficiency of the computation of clutter metrics, as images with near homogeneous clutter level will require only local extent to be computed. Other images may require computation for a few instances only in regions of homogeneous clutter level;
- Gain a better understanding of human vision.

Experimental Procedure

The basic procedure in the experiment is as follows:

- Once started, the experimental program will prompt you for a 3 letter identifier and, once entered, will prompt for a '0' or '1'. A '0' is entered for the first session, (creates your files) and a '1' for each subsequent session (updates your files).
- A monochrome visual stimulus will now be displayed on the video monitor. This consists of two regions, a constant grey level background with a circular textured region in the centre of the background. On 50% of the time, there will be a circular target at the centre of the textured region.
- Your task is to decide if there is in fact a target, which varies in size and contrast, at the centre of the display. Once you have decided hit any key (space bar is convenient) and a prompt menu will appear, which will require you to enter an integer between 1 & 5 indicating a confidence rating in your decision. This will be explained during the training session.
- This will continue until you quit or the last stimulus has occurred. To quit the session hit the 'q' key instead of a number when the prompt screen appears.

Note, the time taken for you to respond to the stimulus will be recorded; *i.e.* the time from when the image appears until you hit the space-bar (or any key).

Training Session(s)

- Type "sizes" and hit the enter key to start a program which displays sequentially the targets in order of size. They are high contrast to enable you to become familiar with the range of target sizes. Each target is displayed for 5 seconds unless a key is hit, which will display the next target. This will continue cycling until you hit the 'q' key.
- To start an actual training session type "training" and hit the enter key. This is similar to the actual experiment, except that feedback is given as to whether a target exists in the image and the clutter background is a constant size (it varies in the real experiment). The program will prompt you for a three letter identifier and, once entered, will prompt for a '0' or '1'. In this training mode it does not really matter, but a '0' is entered for the first session and a '1' for each subsequent session. In the real experiment this is important.

Experimental sessions

- To start the experiment type "go" and hit the enter key. This will present the same prompts as in the training session, but make sure you type '0' for the first session (after the training session) and '1' thereafter.
- Keep the maximum time of each session to 30 minutes, less if tired *etc.*
- To end a session type 'q' instead of a number after an image presentation. This will save your current status and if you type a '1' at the start of the next session start, the experiment will continue from where you left off.

- If you need to pause, you can wait as long as you like once the prompt screen appears, before you enter your confidence rating.

Note, sessions are done at the same time of the day for each subject, but they do not have to be done on consecutive days; *i.e.* A day can be missed here and there, but this will prolong the experiment.

Appendix I

Linear Digital Filtering

Consider a time sequence of length N, $\{x_1, x_2, \ldots x_N\}$. Apply this to a linear filter with impulse response $\{h_1, h_2, \ldots h_M\}$, having output $\{y_1, y_2, \ldots y_{N+M-1}\}$.
y_n is given by a convolution sum. *i.e.*

$$\begin{aligned} y_1 &= \sum_{r=1}^{M} h_r x_{2-r} = h_1 \\ y_2 &= \sum_{r=1}^{M} h_r x_{3-r} = h_2 \\ y_3 &= \sum_{r=1}^{M} h_r x_{4-r} = h_3 \\ &\vdots \\ y_M &= \sum_{r=1}^{M} h_r x_{M+1-r} = h_M \end{aligned} \tag{I.1}$$

i.e. h_r is the "impulse response". It can be shown that

$$\sum_{n=1}^{N} x_n z^{-n} \cdot \sum_{m=1}^{M} h_m z^{-m} = \sum_{r=1}^{N+M-1} (\sum_{m=1}^{M} h_m x_{n+1-m}) z^{-r} \tag{I.2}$$

which is the z-transform form of the convolution theorem. If we define

$$\begin{aligned} X(z) &= \sum_{n=1}^{M} x_n z^{-n} \\ \text{and } H(z) &= \sum_{m=1}^{M} h_m z^{-m} \\ \text{and } Y(z) &= \sum_{r=1}^{N+M-1} y_r z^{-r}. \end{aligned} \tag{I.3}$$

This defines respectively, the z-transforms of $\{x_n\}, \{h_m\} and \{y_r\}$ and this results in

$$Y(z) = H(z) \cdot X(z). \tag{I.4}$$

Suppose that a system to be considered has bounded inputs and bounded outputs. *i.e.*

$$\sum |h_r| < \infty \tag{I.5}$$

$\Rightarrow H(z)$ has all its poles inside $|z| = 1$ *i.e.* the unit circle. Suppose the input is $x_n = e^{j2\pi fn}$ for $-\infty < n < \infty$, then

$$\begin{aligned} y_n &= \sum_{r=-\infty}^{\infty} h_r e^{-j2\pi f(n+1-r)} && \text{(I.6)} \\ &= x_{n+1} \sum_{r=-\infty}^{\infty} e^{-j2\pi fr} && \text{(I.7)} \\ &= x_{n+1} H(e^{-j2\pi f}) && \text{(I.8)} \end{aligned}$$

where $H(e^{-j2\pi f})$ is the frequency response of the system. Define $R_{yy}(k) = E[y_n y^*_{n+k}]$ as the auto-covariance function (acvf) of the output of the filter, and the spectral density $S_{yy}(f) = T_s \sum_{k=-\infty}^{\infty} R_{yy}(k) e^{-j2\pi jkT_s}$, where T_s is the sampling interval.
Now

$$\begin{aligned} R_{yy}(k) &= E[(\sum_{r=1}^{M} h_r x_{n+1-r})(\sum_{s=1}^{M} h^*_s x_{n+1-s})] && \text{(I.9)} \\ &= \sum_{r=1}^{M} \sum_{s=1}^{M} h_r h^*_s E[x_{(n+1-r)} x^*_{(n+1+k-s)}] && \text{(I.10)} \\ &= \sum_{r=1}^{M} \sum_{s=1}^{M} h_r h^*_s R_{xx}(k+r-s) && \text{(I.11)} \end{aligned}$$

Then

$$\begin{aligned} S_{yy}(f) &= \sum_{r=1}^{M} \sum_{s=1}^{M} h_r h^*_s \sum_{k=-\infty}^{\infty} R_{xx}(k+r-s) e^{-j2\pi fkT_s} && \text{(I.12)} \\ &= \sum_{r=1}^{M} h_r e^{-j2\pi frT_s} \sum_{s=1}^{M} h^*_s e^{j2\pi fsT_s} && \text{(I.13)} \\ &= H(z) H^*(z) S_{xx}(f) && \text{(I.14)} \\ &= S_{xx}(f) |H(z)|^2 && \text{(I.15)} \end{aligned}$$

Appendix J

Summary of Experimental Procedure for Chapter 9

The following is a guide to carrying out the preparation for this experiment. Each step is briefly described. Steps that are optional (that provide auxiliary information) are marked thus: †.

- Preparing the imagery.
  - (i) Extract targets from spot images.
    - (a) Extract tarmac areas manually.
    - (b) Extract targets.
    - (c) † Make montage images of the extracted targets.
  - (ii) Prepare background images.
    - (a) Choose sufficient imagery in gips cell format with radiometric correction applied.
    - (b) Chop it into approximately 150 megapixel pieces, rounded to block size (64×64).
    - (c) Classify each block according to clutter.
      - i. Measure clutter and average brightness of each block.
      - ii. Histogram the clutter and brightness for all blocks.
      - iii. Manually set a clutter and brightness threshold to remove shadow regions.
      - iv. Divide remainder of clutter histogram into three equal areas.
      - v. Make lists of image blocks belonging to different clutter regimes.
    - (d) Find existing targets in the images.
  - (iii) Insert targets into background images.
    - (a) Mark extended regions the analysts should ignore.
    - (b) Choose positions for targets with manual confirmation.
    - (c) Insert the targets into the imagery at chosen locations and contrast.
- Running the experiment.

(i) Construct directory structure to contain results and programs for running the experiment.

(ii) Prepare named directories and edit the list of analysts in `roc-choose`.

(iii) Prepare account for running the experiment with X and `roc-choose` started on login.

(iv) Prepare timetable of daily session times for analysts.

(v) Instruct each analyst with the training image.

(vi) Analysts perform the task.

(vii) † Check how much each analyst has done.

(viii) † Show basic statistics on each analyst's detections.

- Preparing data for analysis.

(i) Extract all potential detections from `displaytool`'s bizarre log format into a list of unique detections.

(ii) † Make montage image of the detections.

(iii) Measure clutter and contrast of all detections.

(iv) Manually classify each unique detection.

(v) Collate the measurements on detections into one table.

(vi) Produce data in format suitable for producing ROC.

Appendix K

Questionnaire Used in Study Detailed in Chapter 9

The text of the questionnaire is included here, followed by the analysts' lightly edited responses.

Questionnaire for Analyst Observers

Installation: Man-made object that is a permanent piece of infrastructure. Determined from other associated or adjacent infrastructure *i.e.* road signs occurring at intersections, power poles occurring at regular intervals.

1. What did you do when you encountered installations?
   a. Ignored it.
   b. Logged it with a particular confidence (what value of confidence?).
   c. Logged it with a confidence in proportion to its brightness.
   d. Other. (please explain)

Responses:

Subject 1 c.

Subject 2 If it was an obvious installation with no 4WD sized objects then a. Otherwise I tended to log 4WD like installations (*e.g.*, poles, *etc.*) as an "unsure".

Subject 3 c.

Subject 4 c.

Subject 5 b. I usually logged it with a high confidence knowing that it was probably a man-made object (although possibly *not* a 4WD).

Subject 6 c.

Subject 7 If the installation wasn't "boxed" I logged the obvious bright targets with confidence 4 (I think), less bright targets with 3 (I think). Sometimes installations were "boxed". However, the box didn't encompass all targets within the broad vicinity. I generally logged such brights lying outside the "boxed" area as targets as well. I

only ignored targets inside a boxed area and targets *just* outside the boxed area (if there were any).

Installations were useful brightness calibrators. After looking at dozens of screens you start seeing "ghosts". Seeing an installation allows you to identify what a bright target actually is. In my first attempt (*i.e.* first ROC experiment) I logged lots of targets to start off with as I really didn't know what a target looked like.

Subject 8 c. I generally gave low confidences to these

Subject 9 I ignored all objects that I thought were power poles or street signs.

Subject 10 a.

2. What did you do when you encountered an object that was not an installation (in your estimation) but was also not a 4WD?
   a. Ignored it.
   b. Logged it with a particular confidence (what value of confidence?).
   c. Logged it with a confidence in proportion to its brightness.
   d. Other. (please explain)

Responses:

Subject 1 c.

Subject 2 If it was something that could potentially be a 4WD, (*e.g.*, a radar bright in a clearing that didn't clearly look like a tree), then it usually got a "unsure". A lot of these "unsure" detections were features that in an operational scenario I would "follow up on" in terms of looking at a zoomed version, different stretch, previous data sets, other geographical information, *etc.* Prior to the change in the detection guidelines I would "unsure" any feature I would want to examine in more detail, including wanting only to do a zoom or stretch.

Features that appeared more 4WD like received a higher rating. Some features would sometimes get the higher rating by mistake if I forgot to change the confidence. In several cases I would add a second detection to the same point with a lower confidence.

Subject 3 c.

Subject 4 c.

Subject 5 b. (as above).

Subject 6 a. (or b. with confidence 1).

Subject 7 As I have no experience in distinguishing target types I generally logged targets on a brightness scale. Assuming a bright targets were 4WD I generally logged them as confidence 3. Generally less bright targets were 2. Targets of low brightness and difficult to see but which looked out of place were a 1.

Subject 8 a.

Subject 9 I logged all non-installation objects with a confidence proportional to their brightness. I have absolutely no idea what a 4WD looks like and I also had no idea what the pixel sizes represented in terms of scale.

Subject 10 a.

3. What clues did you use in detecting a target? Please explain in your own words.

Responses:

Subject 1 General background and context (*i.e.* trees, grass, near road, near other brights, size, brightness).

Subject 2 Radar brightness, size, presence of a clear radar shadow and this shadow's shape, difference to nearby objects/trees.

Subject 3 For the first part (almost half) of this experiment I looked for objects that appeared artificial. I thus looked at an object's brightness, size and its relationship to its surroundings. In effect I was detecting man-made objects, not trying to determine if a suspected man-made object was a 4WD.

Subject 4 If it was unusually bright compared to the adjacent background.

Subject 5 Brightness and location/context, although since we knew that the targets had been inserted relatively randomly then location was something I generally tried to ignore. The biggest problem was the "unsures", as I know that if I was reporting to someone (*a la* K95) then I wouldn't have reported these.

Subject 6 Brightness with respect to the average background was my major criterion. If there was a lot of clutter, my subjective brightness threshold would go up, and vice versa. I discriminated against brights on ridge lines, especially if there was an arc of brights from rock lines. In doubtful cases the presence of a road or track influenced my decision (positively).

Subject 7 As mentioned above, installations helped me calibrate for brightness. Also would generally look harder in the vicinity of installation or roads for targets as one might expect more targets to be around these areas.

Also tried to identify things which looked out of place against the background imagery.

In areas with lots of trees or very patchy one could either see lots of targets or see none at all depending on how thorough you wanted to be and how many targets you expected to see in any given ROC experiment.

Subject 8 I looked to see if there were any obvious areas that were brighter than the surrounding landscape. I also considered the size and shape of the suspect area to determine if I thought it was a target. Looked for shadows, trees should have longer shadows than 4WD since they are taller.

Subject 9 Unusually bright objects with respect to their immediate surroundings.

Subject 10 Isolated brights, not too large, not near a large number of similar bright spots, not on water, not regularly spaced.

4. What search strategy did you use?

Responses:

Subject 1 Across and down the image one screenful at a time.

Subject 2 Quick overview of the screen, focus on the radar bright/high contrast features, then look for features different to the natural clutter, final check of the image near the screen edges because I found myself occasionally focusing more on the centre.

Subject 3 I scanned each "screen view" from left to right working from top to bottom.

Subject 4 Generally I zig-zagged across the image one screen at a time, but occasionally I stepped back one screen to recheck something.

Subject 5 I generally used a raster scan. However, I also on occasion used a circular scan *i.e.* top left to top right to bottom right to bottom left. I found that you had to be careful to make sure that you viewed all the screen equally.

Subject 6 First a rough, quick scan over the image, to note anything that was bright and could be a 4WD, with a special look in corners and near the boundaries and along tracks. I then looked more thoroughly at what I had mentally noted to make a decision. I did not agonise over uncertain ones, especially in high clutter images.

Subject 7 Principally looked for bright targets, especially near roads or installations. Or looked for targets that seemed out of place.

Subject 8 I started looking in the top left corner of the screen and just scanned across and down each and across and down *etc.* each screen.

Subject 9 I usually spent about 10 seconds per screen. Generally I would break the screen up into about 5 horizontal regions and search from left to right, drop down to the next region, go from right to left and so on.

Subject 10 Frame to frame as suggested, within a frame quick scan over whole image. If many bright spots or no spots over whole frame then quickly move on. If isolated spots seen then more detailed examination of likely looking points.

5. Is it possible that the room lights were not on at some time?

Responses:

Subject 1 No.

Subject 2 No.

Subject 3 Yes.

Subject 4 The lights were on each time.

Subject 5 No.

Subject 6 The lights were definitely on at all times.

Subject 7 I'm pretty sure the lights were on all the time.

Subject 8 Yes, on one occasion at least I remember someone working in the room too, and as such the lights were on.

Subject 9 No.

Subject 10 Sometimes lights not on at start but I'm fairly sure I remembered to turn them on.

Appendix L

Instructions to Observers in Study Detailed in Chapter 9

Startup

- Login as *ingara* on basil if necessary. You should be automatically using the "openwindows" display manager.
- Select your name from the menu and click "OK".
- "DISPLAYTOOL" will appear slowly with your next image preloaded.
- You can check your progress on the "overview" window at any time by toggling the full-screen zoom using "Open" key on the left of your keyboard. Remember though that the clock is recording all your time, and this will be important in the analysis.

Searching

- Carefully examine the image before you for targets.
- If you spot one (or more), then make a detection using the steps described below.
- When you have finished examining the image before you, move to the right using the cursor keys. (Don't use half-pan.) Proceed through the image in a zig-zag fashion, *i.e.* when you can't move any further to the right move down one step and proceed along to the left, and so on.
- You will find targets in the imagery that have been placed there artificially as well as man-made objects present in the scene. Some will obviously be cultural features that are not targets, *e.g.*, a regular line of brights next to a road would indicate electricity poles, and you don't need to log these, but log *everything* else that you think might be, along with your confidence.

- Towns and other extended areas in the original images contain lots of cultural features that are not of interest and should therefore be ignored. To help you know where these locations are precisely, the areas have been bounded by a bright rectangular outline, sometimes with cross-hatching.

Making a Detection

- Detections are made by pressing ctrl-left-mouse with the cursor placed over the target.
- A dialog pops up with three fields. Only the *confidence* field is important. By default, this field will be highlighted so all you need to do is hit a key 1–5 and then press "return". The dialog will disappear.
- The confidence values range from 1 (unsure), 2 (maybe), 3 (likely), 4 (more than likely) and 5 (sure).
- Please try and use *all* confidence values in a systematic way. They correspond to points on the ROC curve.

Exiting

- Don't stop in the middle of a session, only when you have completed an image. Half an hour should be plenty of time for this. The clock starts when the image is presented and stops when you quit.
- Ignore all interruptions — your time is being recorded as a measure of target detectability.
- When finished, exit a session by clicking on the "File" menu, "Close" item at the top-left of the zoom window. This ensures that the detection data is written out to file and stops the clock.

General Procedures and Restrictions

- Please don't change the brightness of the image either by fiddling with the monitor controls or DisplayTool's contrast stretch. They have been set to particular calibrated values.
- Don't use the amplitude threshold either.
- Keep the lights on.
- Don't use level 2 zoom.
- Don't change the zoom factor from "1" on the zoom window, as this will bias the results and cause the detections to be logged incorrectly.

- Don't change the size of the zoom window except by iconizing it using the "open" key for quick looks at the "overview" window.
- Log all targets separately. Don't just log the middle of a group.
- Leave the mouse up the top when you're scanning the image so you don't obscure part of it.

Scenario

Imagine that you are an NCO whose task is to search through *all* the imagery in an efficient but thorough manner to find all targets of military interest that they might contain, except in the hatched regions. These targets of military interest are vehicles, because we are looking for evidence of small mobile forces in the area. Remember, it will be your assessment of the situation that will determine if a helicopter, or some smaller response, should be sent out to investigate. The confidence values range from 1 (unsure), 2 (maybe), 3 (likely), 4 (more than likely) and 5 (sure). When you make a detection of 1, "unsure", the cost of this detection in terms of reconnaissance assets is low — a couple of rookies will be sent out with a pair of binoculars. However, if you make a "certain" detection of 5, the helicopter gunship will investigate and it will be prepared for trouble. If they only find an old oil drum, then they won't be too upset, and everyone will chalk it up to local knowledge. However, if they only find a particularly tall gum tree, then your name will be mud round the base.

You are not interested in noting buildings and power poles, so if you can tell them apart to some degree, then indicate this using the confidence measure. For example, you note a bright object close to a runway. It seems to be bigger than what you would expect for a 4WD, and its co-location with the runway indicates to you that it is a building. If you are confident of this you do not make a detection and you move on. Or, you might not be totally sure of this assessment, so you make it an "unsure" detection with a detection confidence of 1.

Later, you notice a line of brights regularly spaced next to a road — you are confident that these are power poles, so you ignore them and move on. You notice a bright not far away, near a track on flat ground — it looks suspicious. Is it a bright tree? No, you don't think so, because it doesn't have any shadow on the same side as all the other similar textures in the region. It's the right shape and size — you think all these things together give you a reasonable degree of confidence that it is a target of military interest, so make a confidence 4 detection.

Detections

In the images 17, 18 and 19 you will find some regions of high activity. We want you to log all the points in there that you think are not permanent installations, *i.e.* vehicles, along with how confident you are. We will be examining every detection that everyone makes and we are including in our analysis targets that we didn't place in the image as well as those that we did.

The moral? Log *everything* suspicious in the image along with your confidence outside the no-go regions, even if you think we didn't put it there.

Please note that you cannot change the confidence of a detection once you hit the "ok" button. If you have made a mistake, then as a last resort you can make a second detection at the same spot with the correct confidence. We would discourage too much of this though.

Environment

Please leave the lights on during the experiment. We will be measuring the luminance of the room and of the screen at all grey levels, so if you turn the lights out you will bias your results and possibly ruin them.

Also, please refrain from peeking at the results. This also could bias the results. Don't worry how others are doing — it is not a competition and in the end we will be averaging them across all observers. Work at a speed that is comfortable for you. It doesn't matter if you need longer, as long as you get through all the images in the end.

General

- Just a point of clarification for everyone: when you are practised, if you wish, you are entitled to do more than one image in a session, as long as you do not sit in front of the machine for more than roughly half an hour. Please take the experiment seriously and we won't have to do it again.
- We will be making the data anonymous, in case anyone is concerned about confidentiality. The results will be used to compute an average measure.
- Once you begin the experiment, it is vital that you complete the whole series, or your hard work will not be useful to us.
- You will be timed, so please ignore any interruptions during this time.
- We want people to stick as closely as possible to the time slot allotted.

Bibliography

Abdou, I. E. and Pratt, W. K. (1979). Quantitative design and evaluation of enhancement/thresholding edge detector. *Proceedings of the IEEE*, 67(5):753–763.

Ahmed, N., Natarajan, T., and Rao, K. (1974). Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93.

Ajjimarangsee, P. and Huntsberger, T. L. (1988). Neural network model for fusion of visible and infrared sensor outputs. In Schenker, P. S., editor, *Proceedings of the SPIE on Sensor Fusion:Spatial Reasoning and Scene Interpretation*, volume 1003, pages 153–160.

Algazi, V. R. and Ford, G. E. (1980). Quality measures in the processing of high contrast images. In Tescher, A. G., editor, *Proceedings of the SPIE on Advances in Image Transmission II*, volume 249, pages 54–60.

ANSI/HFS (1988). American national standard for human factors engineering of visual display terminal workstations. Joint American National Standards Institute and the Human Factors Society Standard Number: 100-1988.

Arps, R. (1979). Binary image compression. In Pratt, W., editor, *Image Transmission Techniques*, pages 219–274. Academic Press, New York.

Avadhanam, N. and Algazi, V. (1996). Prediction and measurement of high quality in still image coding. In Algazi, V., Ono, S., and Tescher, A. G., editors, *Proceedings of the SPIE on Very High Resolution and Quality Imaging*, volume 2663, pages 100–109.

Barlow, H. (1958). Temporal and spatial summation in human vision at different background intensities. *Journal of Physiology*, 141:337–350.

Barrett, H. H., Yao, J., Rolland, J. P., and Myers, K. J. (1993). Model observers for assessment of image quality. In *Proceedings of the National Academy of Sciences (USA)*, volume 90, pages 9758–9765.

Barten, P. (1987). The sqri method:a new method for the evaluation of visible resolution on a display. *Journal of the Society for Information Display*, 28(3):253–262.

Bartleson, C. (1975). Optimum image reproduction. *Journal of the Society of Motion Picture & Television Engineers*, 84(8):613–618.

Bartleson, C. and Breneman, E. (1967). Brightness reproduction in the photographic process. *Photographic Science & Engineering*, 11:254–263.

Beckenbach, E. F. and Bellman, R. (1971). *Inequalities*. Springer-Verlag, New York.

Bedworth, M. D. and Bridle, J. S. (1987). Experiments with the back-propagation algorithm: A systematic look at a small problem. Technical report, Royal Signals and Radar Establishment, Malvern, UK.

Berger, T. (1972). Rate distortion theory for context dependent fidelity. *IEEE Transactions on Information Theory*, IT-18(3):378–384.

Bertilone, D. C., Caprari, R. S., Angeli, S., and Newsam, G. N. (1997). Spatial statistics of natural terrain imagery. i. non-gaussian ir backgrounds and long-range correlation. *Applied Optics*, 36:9167–9176.

Bertilone, D. C., Caprari, R. S., Chapple, P. B., and Angeli, S. (1998). Spatial statistics of natural terrain imagery. ii. oblique visible backgrounds and stochastic simulatio. *Optical Communication*, 150:71–76.

Biberman, L. (1973). *Perception of Displayed Information*. Plenum Press.

Blackwell, R. H. (1946). Contrast thresholds of the human eye. *Journal of the Optical Society of America*, 36(11):624–643.

Briggs, S. (1980). The definition and measurement of image quality. In Tescher, A. G., editor, *Proceedings of the SPIE on Advances in Image Transmission II*, volume 249, pages 170–174.

Brown, D. and Ballard, C. (1982). *Computer vision*. Prentice-Hall ,Englewood Cliffs.

Budrikis, Z. (1972). Visual fidelity criterion and modeling. *Proceedings of the IEEE*, 60(7):771–779.

Buffett, A. (1986). Visual perception of visual contrast:implications for the assessment of visual quality. *Displays-Technology & Applications*, 17(1):30–36.

Bunch, P. C., Hamilton, J. F., Sanderson, G. K., and Simmons, A. H. (1978). A free response approach to the measurement and characterization of radiographic observer performance. *Journal of Applied Photographic Engineering*, 4(4):166–171.

Burgess, A. E. (1989). On sampling statistics in observer performance studies. In Viergever, M. A., editor, *Proceedings of the SPIE on Science and Engineering of Medical Imaging*, volume 1137, pages 190–197.

Burgess, A. E. (1995). Comparison of non-whitening and hotelling observer models. In Kundel, H. L., editor, *Proceedings of the SPIE on Medical Imaging 1995: Image Perception*, volume 2436, pages 2–9.

Burr, D. C., Morrone, M. C., and Ross, J. (1979). Local and global visual processing. *Vision Research*, 26(5):749–757.

Burt, P. and Adelson, E. (1983). The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, COM-31(4):532–540.

Caelli, T. and Julesz, B. (1979). Psychophysical evidence for global processing in visual texture discrimination. *Journal of the Optical Society of America*, 69(5):675–678.

Cathcart, M., Doll, T., and Schmeider, D. (1989). Target detection in urban clutter. *IEEE Transactions on Systems Man & Cybernetics*, SMC-19(5):1242–1250.

Chapple, P. (1997). Personal Communication.

Chen, C., Len, D. W., and Hsing, T. R. (1995). Digital visual communications over telephone networks. *Journal of Visual Communication & Image Representation*, 6(2):97–108.

Chen, D. and Bovik, A. (1990). Visual pattern image coding. *IEEE Transactions on Communications*, 38(12):2137–2145.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

De Ridder, H. and Majoor, G. (1990). Numerical category scaling: An efficient method for assessing digital image coding impairments. In Allebach, J. P. and Rogowitz, B. E., editors, *Proceedings of SPIE on Human Vision and Electronic Imaging:Models, Methods and Applications*, volume 1249, pages 65–77.

Dodd, N. (1986). Texture generation using fractal concepts. In *Second International Conference On Image Processing and its Application*, pages 253–257. The Institute of Electrical Engineers (UK).

Doll, T. and Schmieder, D. (1993). Observer false alarm rates on detection in clutter. *Optical Engineering*, 32(7):1675–1684.

Doll, T. J., McWhorter, S. W., and Schmeider, D. E. (1993). Target and background characterization based on a simulation of human perception. In Watkins, W. R., editor, *Proceedings of the SPIE on Characterization, Propogation and Simulation of Sources and Backgrounds III*, volume 1967, pages 432–454.

Dom, B. (1981). Gray scale compression in reprography. *Photographic Science & Engineering*, 25(5):216–218.

Egan, J. P. (1975a). *Signal Detection Theory and ROC Analysis*, chapter 2. Academic Press. Section 2.6.

Egan, J. P. (1975b). *Signal Detection Theory and ROC Analysis*, chapter 6. Academic Press.

Egan, J. P. (1975c). *Signal Detection Theory and ROC Analysis*. Academic Press.

Eggerton, J. and Srinath, M. (1986). A visually weighted scheme for image bandwidth compression at low data rates. *IEEE Transactions on Commmunications*, 34(8):840–847.

Evans, S. and Attaya, W. L. (1978). Research methods and strategies in the psychophysics of image quality. *Photographic Science & Engineering*, 22(2):92–97.

Ewing, G. (1991). Image quality measures.

Ewing, G. J. and Woodruff, C. J. (1996). Comparison of jpeg and fractal based image compression on target acquisition by human observers. *Optical Engineering*, vol 35(1):284–288.

Falmagne, J. (1986). Psychophysical measurement theory. In K. Boff, L. K. and Thomas, J., editors, *Handbook of Perception and Human Performance*, volume 1. John Wiley & Sons, New York, first edition.

Fechner, G. (1966). *Elements of Psychophysics*. Holt, Rinehart and Winston. Originally published 1860.

Feng, Y., Ostberg, O., and Lindstrom, B. (1990). Mfta as a measure for computer display screen image quality. *Displays-Technology & Applications*, 11(4):186–192.

Fisher, Y. (1993). Fractal image compression. *Fractals*, 2(3):347–361.

Fitzsimons, B. (1998). Synthetic aperture radar: Pictures from radio waves. *Armada International*, (4):28–35.

Fuhrmann, D., Baro, J., and Cox, J. (1995). Experimental evaluation of psychophysical distortion metrics for jpeg encoded images. *Journal of Electronic Imaging*, 4(4):397–405.

Gaito, J. (1960). Scale classification and statistics. *Psychological Bulletin*, 67:277–278.

Gendron, R. (1973). An improved objective method for rating picture sharpness: Cmt acutance. *Journal of the Society of Motion Picture & Television Engineers*, 82(12):1009–1012.

Gonzalez, R. and Wintz, P. (1987a). *Digital Image Processing.*, chapter 8, pages 428–448. Addisson-Wiley.

Gonzalez, R. and Wintz, P. (1987b). *Digital Image Processing.*, chapter 7, pages 336–340. Addisson-Wiley.

Gonzalez, R. and Wintz, P. (1987c). *Digital Image Processing.*, chapter 5, page 229. Addisson-Wiley.

Goodenough, D. J. (1976). Assessment of image quality of diagnostic imaging systems. In *Medical Images: Formation, Perception and Measurement. Proceedings of the Seventh L. H. Gray Conference*. University of Leeds, The Institute of Physics.

Gotlieb, C. C. and Kreyszig, H. E. (1990). Texture descriptors based on co-occurrence matrices. *Computer Vision, Graphics, & Image Processing*, 51(1):70–86.

Granger, E. M. and Cupery, K. N. (1972). An optical merit function (sqf) which correlates with subjective image judgements. *Photographic Science & Engineering*, 16(3):221–30.

Green, D. M. and Swets, J. A. (1966a). Assumed distribution of signal and noise. In *Signal Detection Theory and Psychophysics*, pages 53–85. Peninsula Publishing, Los Altos California, 1988 reprint edition.

Green, D. M. and Swets, J. A. (1966b). *Signal Detection Theory and Psychophysics.* Peninsula Publishing, Los Altos California, 1988 reprint edition.

Green, D. M. and Swets, J. A. (1966c). *Statistical Decision Theory and Psychophysics.*, chapter Energy-Detection Model for Audition, pages 209–232. Peninsula Publishing, Los Altos California, 1988 reprint edition.

Green, D. M. and Swets, J. A. (1966d). *Statistical Decision Theory and Psychophysics*, chapter Statistical Decision Theory and Psychophysical Procedures, pages 45–50. Peninsula Publishing, Los Altos California, 1988 reprint edition.

Green, M. (1992). Visual search: Detection, identification, and localization. *Perception*, 21:765–777.

Gur, Y. (1985). Fourier analysis of image raggedness. In Granger, E. M., editor, *Proceedings of the SPIE on Image Quality an Overview*, volume 549, pages 22–24.

Hall, C. (1977). Image filtering based on psychovisual characteristics of the human visual system. Technical report, Image Processing Institue of the University of Southern California.

Hall, C. (1981). Subjective evaluation of a perceptual quality metric. In Cheatham, P. S., editor, *Proceedings of the SPIE on Image Quality*, volume 310, pages 200–204.

Hall, C. F. and Hall, E. L. (1977). A non-linear model for the spatial characteristics of the human visual system. *IEEE Transactions on Systems Man & Cybernetics*, SMC-7(3):161–170.

Hamerly, J. R. and Springer, R. M. (1981). Raggedness of edges. *Journal of the Optical Society of America*, 71(3):285–288.

Hamming, R. (1980). *Coding and information theory.* Prentice-Hall, Englewood Cliffs New Jersey.

Hanley, J. A. and McNeil, B. J. (1982). The meaning of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Haralick, R. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–803.

Haralick, R., Shanmugan, K., and Dinstein, T. (1973). Textural features for image classification. *IEEE Transactions on Systems Man & Cybernetics*, SMC-3(6):610–621.

He, D.-C. and Wang, L. (1991). Texture features based on texture spectrum. *Pattern Recognition*, 24(5):391–399.

Herrmann, C., Buhr, E., Hoeschen, D., and Fan, S.-Y. (1993). Comparison of roc and afc methods in a visual detection task. *Medical Physics*, 20(3):805–811.

Hidaka, T. and Ozawa, K. (1993). Iso/iec jtc1 sc29/wg11 and report on mpeg-2 subjective assessment at kurihana. *Signal Processing: Image Communication*, 5:127–157.

Higgins, G. (1977). Image quality criteria. *Journal of Applied Photographic Engineering*, 3(2):53–60.

Hilgers, J. W., Vockel, W. P., and Reynolds, W. R. (1997). Sensor and detection algorithm based clutter metrics. In Watkins, W. R., editor, *Proceedings of the SPIE on Targets and Backgrounds: Characterization and Representation III*, volume 3062, pages 267–277.

Hines, W. W. and Montgomery, D. C. (1980a). Analysis of variance. In *Probablity and Statistics in Engineering and Management Science*, pages 328–352. John Wiley & Sons, second edition.

Hines, W. W. and Montgomery, D. C. (1980b). Design of experiments. In *Probablity and Statistics in Engineering and Management Science*, pages 481–485. John Wiley & Sons, second edition.

Hord, R. (1982). *Digital Image Processing of Remotely Sensed Data.* Academic Press, New York.

Horn, B. K. P. (1986). *Robot Vision.* MIT Press, Cambridge Massachusetts.

Huang, T. (1965). Pcm picture processing. *IEEE Spectrum*, 2(12):57–63.

Infante, C. (1991). Numerical methods for computing modulation transfer-function areas. *Displays-Technology & Applications*, 12(2):80–83.

ISO/IEC (1998). Mpeg-4 overview - (dublin version). Koenen, R. (Ed.).

Jaquin, A. E. (1993). Fractal image coding: A review. *Proceedings of the IEEE*, 81(10):1451–1465.

Julesz, B. (1991). Early vision and focal attention. *Reviews of Modern Physics.*, 63(3):735–72.

Kak, A. and Slaney, M. (1988). *Principles of Computerized Tomographic Imaging.* IEEE Press.

Karbowiak, A. (1969a). *Theory of Communication.* Oliver & Boyd, Edinburgh.

Karbowiak, A. (1969b). *Theory of Communication*, page 44. Oliver & Boyd, Edinburgh.

Kaukoranta, T. and Nevalainen, O. (1996). Empirical study on subjective quality evaluation of compressed image. In Algazi, V., Ono, S., and Tescher, A., editors, *Proceedings of the SPIE on Very High Resolution and Quality Imaging*, volume 2663, pages 88–99.

Keppel, G. (1991a). *Design and Analysis: A Researcher's Handbook*, chapter 7: Analysis of Trend. Prentice Hall.

Keppel, G. (1991b). *Design and Analysis: A Researcher's Handbook*, chapter 13, pages 282–284. Prentice Hall.

Keppel, G. (1991c). *Design and Analysis: A Researcher's Handbook*, chapter 13, pages 284–293. Prentice Hall.

Keppel, G. (1991d). *Design and Analysis: A Researcher's Handbook*. Prentice Hall.

Kingdom, F. and Moulden, B. (1989). Modelling visual detection: Luminance response non-linearity and internal noise. *Quarterly Journal of Experimental Psychology*, 41A:675–696.

Klymenko, V., Pizer, S. M., and Johnstone, R. E. (1990). Visual psychophysics and medical imaging: Nonparametric adaptive method for rapid threshold estimation in sensitivity experiments. *IEEE Transactions on Medical Imaging*, 9(4):353–365.

Knill, D. C., Field, D., and Kersten, D. (1990). Human discrimination of fractal images. *Journal of the Optical Society of America A*, 7(6):1113–1123.

Korovkin, P. (1961). *Inequalities*. Pergamin Press.

Kostas, T., Sullivan, B., Ansari, R., Giger, M., and MacMahon, H. (1993). Adaption and evaluation of jpeg-based compression for radiographic images. In Loew, M. H., editor, *Proceedings of the SPIE on Medical Imaging 1993: Image Capture, Formatting, and Display*, volume 1897, pages 276–281.

Kusaka, H. (1989). Consideration of vision and picture quality - psychological effects induced by picture sharpness. In Rogowitz, B. E., editor, *Proceedings of the SPIE on Human Vision, Visual Processing and Digital Display*, volume 1077, pages 50–55.

Laming, D. (1986). *Sensory Analysis*. Academic Press.

Leger, A., Omachi, T., and Wallace, G. (1991). Jpeg still picture compression algorithm. *Optical Engineering*, 30(7):947–954.

Levine, M. (1985). *Vision in Man and Machine*. McGraw-Hill, New York.

Limb, J. (1979). Distortion criteria of the human viewer. *IEEE Transactions on Systems Man & Cybernetics*, SMC-9(12):778–793.

Lingard, D. M. and Stacy, N. J. S. (1997). A mission model for broad-area aerial surveillance. Technical Report DSTO-TR 0536, DSTO. Classified Report.

Lippmann, R. P. (1987). An introduction to computing with neural nets. *IEEE Acoustics Speech & Signal Processing Magazine*, 4(2):4–22.

Lloyd, J. M. (1975). *Thermal Imaging Systems*. Plenum Press, New York.

Lohscheller, H. (1984). A subjectively adapted image communication system. *IEEE Transactions on Communications*, COM-32(12):1316–1322.

Loo, L., Doi, K., and Metz, C. E. (1984). A comparison of physical image quality indices and observer performance in radiograhic detection of nylon bead. *Physics in Medicine & Biology*, 29(7):837–856.

Lourens, J. (1991). Objective image impairment detection using image processing. In *4th Internal Conference on Television Measurements*, pages 13–19. IEE London.

Lu, G. (1993). Fractal image compression. *Signal Processing:Image Communication*, 5:327–343.

Lukis, F. and Budrikis, Z. L. (1982). Picture quality prediction based on a visual model. *IEEE Transactions on Communications*, Comm-30:1979–1992.

Malo, J., Pons, A., and J.M., A. (1997). Subjective image fidelity metric based on bit allocation of the human visual system in the dct domain. *Image and Image Computing*, 15:535–548.

Mannos, J. L. and Sakrison, D. J. (1974). The effects of a visual fidelity criterion on the encoding of image. *IEEE Transactions on Information Theory*, IT-20(4):525–536.

Marlow, H. (1958). Temporal and spatial summation in human vision at different background intensities. *Journal of Physiology*, 141:337–350.

Marmolin, H. (1986). Subjective mse measures. *IEEE Transactions on Systems Man & cybernetics*, SMC-16(3):486–489.

Marr, D. (1982). *Vision- A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H. Friedman & Co., New York, N.Y.

Meitzler, T., Jackson, W., Sohn, E., and Bednarz, D. (1993). Calculation of background clutter in infrared imagery: A semi-empirical study. In Watkins, W. R., editor, *Proceedings of the SPIE on Characterization, Propogation and Simulation of Sources and Backgrounds III*, volume 1967, pages 525–532.

Meitzler, T. J. (1995). *Modern Approaches to the Computation of the Probability of Target Detection in Cluttered Environments.* PhD thesis, Wayne State University, Detroit Michigan, U.S.A.

Metz, C. (1977). An overview of some measures of image quality. In Gray, J. E. and Hendee, W. R., editors, *Proceedings of the SPIE on Optical Instrumentation in Medicine VI*, volume 127, pages 486–489.

Metz, C. (1986). Roc methodology in radiologic imaging. *Investigative Radiology*, 21:720–733.

Metz, C. E. (1978). Basic principles of roc analysis. *Seminars in Nuclear Medicine*, VIII(4):283–298.

Metz, C. E., Starr, S. J., , and Lusted, L. B. (1976). Quantitative evaluation of visual detection performance in medicine: Roc analysis and determination of diagnostic benefit. In *Medical Images: Formation, Perception and Measurement.* The Institute of Physics, John Wiley & Sons.

Mitrinovic, D. (1970). *Analytic Equalities.* Springer-Verlag, Berlin.

Mosteller, F. (1951). Remarks on the method of paired comparisons iii : A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika.*, 16:207–218.

Moulden, B., Kingdom, F., and Gatley, L. (1990). The standard deviation of luminance as a metric for contrast in random-dot images. *Perception*, 19:79–101.

Nasrabadi, N. and King, A. (1988). Image coding using vector quantization: A review. *IEEE Transactions on Communications*, 36(8):535–549.

Newell, A. (1990). *Unified Theories of Cognition.* William James Lectures 1987. Harvard University Press (Cambridge, Mass).

Newsam, G. and Woodruff, C. (1991). Fractal random fields and their use in vision experiments. In *DICTA-91 Digital Image Computing: Techniques and Applications.*, pages 365–372. Australian Pattern Recognition Society.

Newsam, G. N. (1993). Smooth zooming of images. In *Proceedings of the First New Zealand Conference on Image and Vision Computing*, pages 221–228, Auckland, N.Z.

Nill, N. (1985). A visual model weighted cosine transform for image compression and quality assessment. *IEEE Transactions on Communications*, 33(6):551–557.

Ohtsuka, S., Inoue, M., and Watanabe, K. (1988). Quality evaluation of pictures with multiple impairments based on visually weighted error. *Society for Information Display International Symposium. Digest of Technical Papers*, First Edition:428–431.

Ohtsuka, S. and Makoto, K. (1991). Quality evaluation of locally impaired pictures. *Proceedings of the Society for Information Display*, 32(1):19–24.

O'Kane, B., Walters, C., and Agostino, J. (1993). Report on perception experiments in support of thermal performance models. Technical report, U.S. Army Night Vision Laboratory, Fort Belvoir, VA, USA.

Oliver, C. J. and Quegan, S. (1998). *Understanding Synthetic Aperture Radar Images.* Artech House, Boston.

Overington, I. (1974). Visual efficiency- a means of bridging the gap between subjective and objective image quality. In Dutton, D., editor, *Proceedings of SPIE on Image Assessment and Specification*, volume 46, pages 93–102.

Overington, I. (1976a). *Vision and Aquisition.* Pentech Press, London.

Overington, I. (1976b). *Vision and Aquisition*, chapter 8: Rudimentary Search Modelling, pages 164–174. Pentech Press, London.

Overington, I. (1976c). *Vision and Aquisition*, chapter 4: Basic Thresholds for Detection, page 54. Pentech Press, London.

Overington, I. (1976d). *Vision and Aquisition*, chapter 13: Background Structure, page 236. Pentech Press, London.

Overington, I. (1982). Towards a complete model of photopic visual threshold. *Optical Engineering*, 21(1):002–013.

Panda, D. P. and Kak, A. C. (1976). Image restoration and enhancement. Technical report, Purdue University.

Peatman, J. G. (1964a). Analysis of variance. In *Introduction to Applied Statistics*, pages 319–358. Harper and Row, New York.

Peatman, J. G. (1964b). Wilcoxon rank test. In *Introduction to Applied Statistics*, pages 372–375. Harper and Row, New York.

Peli, E., editor (1995). *Vision Models for Target Detection and Recognition.* Information Display. World Scientific.

Pennebaker, W. and Mitchell, J., editors (1993). *JPEG Still Image Data Compression Standard.* Van Nostrand, New York.

Pentland, A. and Horowitz, B. (1991). A practical approach to fractal-based image compression. In Tzou, K., editor, *Proceedings of SPIE on Visual Processing and Image Processing '91: Visual Communication*, volume 1605, pages 467–474.

Pentland, A. P. (1984). Fractal-based image description of natural scenes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, PAMI-6(6):661–673.

Piper, S. (1983). The evaluation of the accuracy of computer classification. *Proceedings of the 9th International Symposium on Machine Processing of Remotely Sensed Data*, pages 303–310.

Pratt, W. (1978). *Digital Image Processing.* Wiley, New York.

Pratt, W. (1979). *Image Transmission Techniques.* Academic Press, New York.

Quincy, E. (1990). Video quality gradient measures for digital networks. *MILCOM 90. A New Era. 1990 IEEE Military Communications Conference*, 1:289–96.

Reynolds, W. (1990). Towards quantifying infrared clutter. In Watkins, W. R., editor, *Proceedings of SPIE on Characterization, Propogation and Simulation of Infrared Scenes*, volume 1311, pages 232–240.

Richards, J. (1986). *Remote Sensing Digital Image Analysis.* Springer-Verlag, Berlin.

Rockette, H. E., Obuchowski, N. A., Gur, D., and Good, W. F. (1991). Effect of experimental design on sample size. In Gilbert, J. R., editor, *Proceedings of SPIE on Medical Imaging V: PACS Design and Evaluation*, volume 1446, pages 276–283.

Rogowitz, B. (1985). A psychological approach to image quality. In Granger, E. M., editor, *Proceedings of SPIE on Image Quality: An Overview*, volume 549, pages 9–13.

Rose, A. (1973). *Vision:Human and Electronic*. Plenum Press, New York.

Rosenfeld, A. and Kak, A. C. (1982). *Digital Picture Processing*. Academic Press, New York, second edition.

Rotman, S., Gordon, E., and Kowalczyk, M. (1991). Modeling human search and target acquisition performance iii : Target detection in the presence of obscurants. *Optical Engineering*, 30(6).

Rotman, S., Tidhar, G., and Kowalczyk, M. (1994). Clutter metrics for target detection systems. *IEEE Transactions on Aerospace & Electronic Systems*, 30(1):81–90.

Schmieder, D. and Weathersby, M. (1983). Detection performance in clutter with variable resolution. *IEEE Transactions on Aerospace & Electronic Systems*, 19(4):662–630.

Snyder, H. (1973). Image quality and observer performance. In *Perception of Displayed Information*. Plenum Press.

Snyder, H. L. (1989). The ansi/hfs standard for visual display terminals. *Information Display*, 5(5):20–3.

Spaulding, K. and Engeldrum, P. G. (1985). A decision theory approach to tone production. In Granger, E. M., editor, *Proceedings of SPIE on Image Quality:An Overview*, volume 549, pages 14–21.

Sperling and Dosher (1986). Psychophysical measurement theory. In K. Boff, L. K. and Thomas, J., editors, *Handbook of Perception and Human Performance*, volume 1. John Wiley & Sons, New York.

Stacy, N. J. S. and Smith, R. B. (1996). Ingara land surveillance trial plan. Technical Report MRD Discussion Document, Microwave Radar Division, DSTO.

Stevens, S. (1968). Measurement, statistics and the schemapiric view. *Science*, 161:849–856.

Stockham, T. (1972). Image processing in the context of a visual model. *Proceedings of the IEEE*, 60(7):828–842.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science*, 182:990–1000.

Tekalp, A. M. (1995). *Video Compression Standards*, chapter 23, pages 432–456. Prentice-Hall.

Thibos, L. (1989). Image processing by the human eye. In Pearlman, W. A., editor, *Proceedings of SPIE on Visual Communications and Image Processing IV*, volume 1199, pages 1148–1153.

Todd-Pokropek, A. (1976). Image processing in nuclear medicine: An examination of the quest for a measure of clinical quality. In *Medical Images: Formation, Perception and Measurement*. The Institute of Physics, John Wiley & Sons.

Toet, A., Ijspeert, J. K., Waxman, A. M., and Aguilar, M. (1997). Fusion of visible and thermal imagery improves situational awareness. *Displays*, 18:85–95.

Trivedi, M. and Rosenfeld, A. (1989). On making computers see. *IEEE Transactions on Systems Man & Cybernetics*, SMC-19(6):1333–1336.

Trivedi, M. M. and Shirvaiker, M. V. (1993). Quantitative characterization of image clutter: Problem, progress, and promises. In Watkins, W. R., editor, *Proceedings of SPIE on Characterization, Propogation and Simulation of Sources and Backgrounds III*, volume 1967, pages 288–299.

USAF (1980). Minimum resolved object sizes for imagery interpretation. Technical Report AIR STD 101/8A, United States Air Force.

van der Lubbe, J. C. A. (1984a). Image quality criteria with emphasis on criteria for remote sensing imagery. Technical report, Netherlands Agency for Aerospace Programs.

van der Lubbe, J. C. A., Boxma, Y., and Boeker, D. (1984). A generalized class of certainty and information measures. *Information Science*, 32(3):187–215.

van der Schaaf, A. and van Hateran, J. (1996). Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759–2770.

Van Dijk, A. M. and Martens, J.-B. (1997). Subjective quality assessment of compressed images. *Signal Processing*, 58:235–252.

Waldman, G., Wootton, J., Hobson, G., and Luetkemeyer, K. (1988). A normalized clutter measure for images. *Computers Vision & Image Processing*, 42:137–156.

Wallace, G. (1992). The jpeg still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):18–34.

Wang, W., Velasco, F. R. D., Wu, A. W., and Rosenfeld, A. (1981). Relative effectiveness of selected texture primitive statistics for texture discrimination. *IEEE Transactions on Systems Man & Cybernetics*, SMC-11(5):360–370.

Watson, A. B. (1993). Dct quantization matrices visually optimized for individual images. In Allebach, J. P., editor, *Proceedings of SPIE on Human Vision, Visual Processing, and Digital Display IV*, volume 1913, pages 202–216.

Wilson, H. R. (1995). Quantitative models for pattern detection and discrimination. In Peli, E., editor, *Vision Models for Target Detection and Recognition*, pages 3–15. World Scientific.

Woodruff, C. J. and Newsam, G. N. (1994). Displaying undersampled imagery. *Optical Engineering*, 33(2):579–585.

Yaglom, A. (1987). *Correlation Theory of Stationary and Related Random Functions.* Springer series in statistics. Springer-Verlag, New York.

Yarbus, A. (1967). *Eye Movements and Vision.* Plenum Press, New York, N.Y.

Young, R. (1986). Simulation of human retinal function with the gaussian derivative model. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 564–569. IEEE Computer Society Press.