Perceptual Vision Models for
Picture Quality Assessment
and Compression Applications

Wilfried Osberger

B.E. (Electronics) (Hons) / B.Inf.Tech
Queensland University of Technology

Space Centre for Satellite Navigation
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
G.P.O. Box 2434, Brisbane, Qld, 4001, Australia

Submitted as a requirement for the degree of Doctor of Philosophy
Queensland University of Technology
March 1999.
To my family
Abstract

The rapid increase in the use of compressed digital images has created a strong need for the development of objective picture quality metrics. No objective metrics have yet been demonstrated to work accurately over a wide range of scenes, coders, and bit-rates. Promising results have been achieved by metrics which model the operation of early stages of the human visual system. However, these models have generally been calibrated using simple visual stimuli, and have not considered any higher level or cognitive processes which are known to influence human perception of image quality.

In this thesis, a new picture quality metric is proposed which models both early vision and higher level attention processes. The early vision model simulates the operation of neurons in the primary visual cortex. The general structure of this model is similar to other models of early visual processes. However, the model’s components have been designed and calibrated specifically for operation with complex natural images. This model is shown to provide accurate prediction of image fidelity. Although the early vision model gives a significantly better prediction of picture quality than simple objective metrics, it is unable to capture any higher level perceptual or cognitive effects. To address this issue, a model of visual attention is developed. This model produces Importance Maps, which predict the regions in the scene that are likely to be the focus of attention. The Importance Maps are used to weight the visible distortions produced by the early vision model, since it is known that distortions occurring in the regions of interest
have the greatest effect on overall picture quality. Comparison with subjective quality rating data demonstrates that the inclusion of the attention model in the quality metric provides a statistically significant increase in the correlation between the model’s prediction and subjective opinion.

Perceptual vision models such as the ones presented in this thesis can be used in many other areas of image processing. Image compression is one of the areas which can obtain significant benefit from the use of perceptual models. To demonstrate this, an MPEG encoder is implemented which uses components of both the early vision and attention models to control the spatially-adaptive quantisation process. Subjective testing of sequences coded using this technique confirms the improvement in subjective quality which can be achieved by using a perceptual vision model in a compression algorithm.
# Contents

Abstract i

Acronyms and Units xiii

Publications xv

Authorship xvii

Acknowledgements xviii

1 Introduction 1

1.1 Assessing the Quality of Compressed Pictures 3

1.2 Scope of the Research 6

1.3 Overview of the Thesis 7

2 The Human Visual System 9

2.1 Optical and Neural Pathways 10

2.2 Properties of Early Vision 20

2.2.1 Sensitivity to Luminance and Contrast Changes 21

2.2.2 Frequency Sensitivities 25

2.2.3 Masking 30

2.3 Eye Movements and Visual Attention 36
2.3.1 Eye Movement Characteristics .......................... 37
2.3.2 Visual Attention ........................................... 41
2.3.3 Relationship Between Eye Movements and Attention ... 47
2.3.4 Factors which Influence Attention ....................... 48

3 Assessing the Quality of Compressed Pictures 58

3.1 Compression Techniques and their Artifacts ............... 61
  3.1.1 Image Compression ........................................ 63
  3.1.2 Video Compression ......................................... 65
  3.1.3 Compression Standards .................................... 66
  3.1.4 Artifacts Introduced by Digital Compression .......... 68

3.2 Subjective Quality Assessment Techniques ................. 70

3.3 Objective Quality Assessment Techniques ................ 75
  3.3.1 PSNR, MSE, and Variants .................................. 76
  3.3.2 Correlation-based Techniques ............................. 77
  3.3.3 HVS-based Techniques .................................... 80
  3.3.4 Other Techniques .......................................... 81

4 Objective Quality Assessment Based on HVS Models 83

4.1 Previous HVS-based Models .................................. 85
  4.1.1 Single Channel Models .................................... 85
  4.1.2 Multiple Channel Models .................................. 87
  4.1.3 Comparison of HVS Models ................................ 93

4.2 Deficiencies of Previous HVS-based Models ................. 96
  4.2.1 Choice of Appropriate Model Parameters .............. 97
  4.2.2 Determining Picture Quality from Visible Distortion Maps 98
  4.2.3 Computational Complexity Issues ....................... 100

4.3 Applying HVS Models to Natural Scenes ................... 101
5 A New Early Vision Model Tuned for Natural Images

5.1 Model Description ............................................. 105
  5.1.1 Inputs to the New Model ................................... 107
  5.1.2 Channel Decomposition and Band-limited Contrast ....... 107
  5.1.3 Contrast Sensitivity ..................................... 110
  5.1.4 Spatial Masking ......................................... 111
  5.1.5 Summation ................................................ 113

5.2 Subjective Image Quality Testing for Model Validation .... 115
  5.2.1 Viewing Conditions ...................................... 115
  5.2.2 Viewing Material ....................................... 116
  5.2.3 Testing Procedure ...................................... 117
  5.2.4 Results of DSCQS Tests ................................ 118

5.3 Performance of the Early Vision Model ....................... 121
  5.3.1 Influence of the Choice of CSF on the Vision Model .... 126
  5.3.2 Influence of Spatial Masking on the Vision Model ...... 128
  5.3.3 Analysis of the New Early Vision Model ............... 128

6 Identifying Important Regions in a Scene ........................ 132

6.1 Previous Computational Models of Visual Attention and Saliency 134
  6.1.1 Multi-resolution-based Attention Models ................ 135
  6.1.2 Region-based Attention Models ......................... 138

6.2 Importance Map Technique for Still Images ................... 140

6.3 Verification of IMs using Eye Movement Data .................. 150
  6.3.1 Description of Eye Movement Experiments .............. 151
  6.3.2 Results of Eye Movement Experiments .................. 152
  6.3.3 Comparison of Fixations and IMs ....................... 152

6.4 Importance Map Technique for Video .......................... 159
B Individual Quality Ratings for DSCQS Test 223

C Fixation Points Across All Subjects 228

Bibliography 235
List of Figures

2.1 Horizontal cross-section of the human eye. .......................... 11
2.2 Density of receptors in the retina as a function of eccentricity. ... 12
2.3 Visual pathway from the eyes to the primary visual cortex. ....... 14
2.4 Cross section of cell layers of the LGN. ............................. 15
2.5 Receptive fields of simple cells. ................................. 16
2.6 Receptive field and possible wiring diagram of a complex cell. ... 17
2.7 Layered structure of the primary visual cortex, showing intercon-
nections between the layers. ............................................ 19
2.8 Modular structure of the primary visual cortex. ................... 20
2.9 Typical luminance change required for detection of stimuli for dif-
ferent background luminance values. ................................. 22
2.10 Typical stimuli used in psychophysical tests to measure contrast. 24
2.11 Typical spatial CSFs at photopic light levels. ..................... 26
2.12 Variation of contrast sensitivity with arbitrary mean luminance
values. ............................................................... 27
2.13 Timing diagram which compares abrupt and gradual temporal pre-
sentation of stimuli. ................................................... 29
2.14 Typical stimuli used in contrast masking tests. ..................... 32
2.15 Threshold elevation caused by contrast masking. .................. 33
2.16 Eye movements for a subject viewing a natural image. .......... 38
2.17 Sample stimuli used in visual search experiments. ............... 44
2.18 Search slopes for efficient and inefficient search. 45
2.19 Sample stimuli used to determine the influence of contrast on visual search. 51
2.20 Sample stimuli used to demonstrate size asymmetry in visual search. 52

3.1 General block diagram of lossy compression algorithms. 62

5.1 Block diagram of the new early vision model. 106
5.2 Structure of filterbanks. 108
5.3 Typical threshold elevation caused by spatial masking. 111
5.4 Sample picture quality rating form used in the DSCQS tests. 117
5.5 Timing of stimulus presentation during DSCQS subjective testing. 118
5.6 Subjective MOS data averaged across all subjects for the first two test images. 119
5.7 Subjective MOS data averaged across all subjects for the final two test images. 120
5.8 Fidelity assessment for the image mandril. 122
5.9 Correlation of objective quality metrics with subjective MOS. 124
5.10 Fidelity assessment for the image lighthouse. 127
5.11 Fidelity assessment for the image football. 129

6.1 Block diagram of the Importance Map algorithm. 141
6.2 IM for the boats image. 147
6.3 IM for the lighthouse image. 149
6.4 Comparison of fixations with IMs for the island image. 154
6.5 The effect of block post-processing on IMs for the island image. 155
6.6 Proportion of hits for each of the test images using the 10% area of highest importance. 156
6.7 Proportion of hits for each of the test images using the 20% area of highest importance. 157
8.8 Comparison scale used for comparing the qualities of the TM5 and perceptually coded sequences. 199

8.9 Comparative subjective quality between the TM5 and perceptually coded sequences (a) Miss America and (b) football. 200

A.1 Test image announcer. (a) Original image, and (b) Importance Map. 212
A.2 Test image beach. (a) Original image, and (b) Importance Map. 213
A.3 Test image boats. (a) Original image, and (b) Importance Map. 214
A.4 Test image claire. (a) Original image, and (b) Importance Map. 215
A.5 Test image football. (a) Original image, and (b) Importance Map. 216
A.6 Test image island. (a) Original image, and (b) Importance Map. 217
A.7 Test image lena. (a) Original image, and (b) Importance Map. 218
A.8 Test image lighthouse. (a) Original image, and (b) Importance Map. 219
A.9 Test image Miss America. (a) Original image, and (b) Importance Map. 220
A.10 Test image pens. (a) Original image, and (b) Importance Map. 221
A.11 Test image splash. (a) Original image, and (b) Importance Map. 222

C.1 Fixations across all subjects for image announcer. 229
C.2 Fixations across all subjects for image beach. 229
C.3 Fixations across all subjects for image boats. 230
C.4 Fixations across all subjects for image claire. 230
C.5 Fixations across all subjects for image football. 231
C.6 Fixations across all subjects for image island. 231
C.7 Fixations across all subjects for image lena. 232
C.8 Fixations across all subjects for image light. 232
C.9 Fixations across all subjects for image Miss America. 233
C.10 Fixations across all subjects for image pens. 233
C.11 Fixations across all subjects for image splash. 234
List of Tables

5.1 Correlation of PQR and MSE with subjective MOS. ............... 123

6.1 Percentage of pixels in each Importance quartile for the 11 test images in Appendix A. ....................... 150

6.2 Proportion of hits across all scenes for different IM block sizes. . 158

7.1 Correlation of IPQR, PQR and MSE with subjective MOS. ....... 176

7.2 Correlation of IPQR with subjective MOS, for different post-processing block sizes. ....................... 177

B.1 MOS for subjects 1–9 for each coded image, following normalisation with respect to the original image. ............... 225

B.2 MOS for subjects 10–18 for each coded image, following normalisation with respect to the original image. ............... 227
# Acronyms and Units

## Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANSI</td>
<td>American National Standards Institute</td>
</tr>
<tr>
<td>CCIR</td>
<td>Comité Consultatif International pour les Radiocommunications</td>
</tr>
<tr>
<td>CSF</td>
<td>Contrast Sensitivity Function</td>
</tr>
<tr>
<td>DCT</td>
<td>Discrete Cosine Transform</td>
</tr>
<tr>
<td>DSCQS</td>
<td>Double Stimulus Continuous Quality Scale</td>
</tr>
<tr>
<td>FIT</td>
<td>Feature Integration Theory</td>
</tr>
<tr>
<td>HDTV</td>
<td>High Definition Television</td>
</tr>
<tr>
<td>HVS</td>
<td>Human Visual System</td>
</tr>
<tr>
<td>IEEE</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>IM</td>
<td>Importance Map</td>
</tr>
<tr>
<td>IOR</td>
<td>Inhibition of Return</td>
</tr>
<tr>
<td>IPDM</td>
<td>Importance Map Weighted Perceptual Distortion Map</td>
</tr>
<tr>
<td>IPQR</td>
<td>Importance Map Weighted Perceptual Quality Rating</td>
</tr>
<tr>
<td>ITU</td>
<td>International Telecommunications Union</td>
</tr>
<tr>
<td>JND</td>
<td>Just Noticeable Difference</td>
</tr>
<tr>
<td>JPEG</td>
<td>Joint Photographic Experts Group</td>
</tr>
<tr>
<td>LBC</td>
<td>Local Band-limited Contrast</td>
</tr>
<tr>
<td>LCD</td>
<td>Liquid Crystal Display</td>
</tr>
<tr>
<td>LGN</td>
<td>Lateral Geniculate Nucleus</td>
</tr>
<tr>
<td>MAD</td>
<td>Mean Absolute Difference</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Score</td>
</tr>
<tr>
<td>MPEG</td>
<td>Moving Picture Experts Group</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>MT</td>
<td>Medial Temporal</td>
</tr>
<tr>
<td>NMSE</td>
<td>Normalised Mean Square Error</td>
</tr>
<tr>
<td>NTSC</td>
<td>National Television System Committee</td>
</tr>
<tr>
<td>PDM</td>
<td>Perceptual Distortion Map</td>
</tr>
<tr>
<td>PQR</td>
<td>Perceptual Quality Rating</td>
</tr>
</tbody>
</table>
PQS  Picture Quality Scale
PSNR  Peak Signal-to-Noise Ratio
QM  Quantisation Matrix
ROG  Ratio of Gaussian
ROI  Region of Interest
RT  Reaction Time
SMPTE  Society of Motion Picture and Television Engineers
SPEM  Smooth Pursuit Eye Movement
SPIE  The International Society for Optical Engineering
TE  Threshold Elevation
TM5  Test Model 5
VDP  Visible Differences Predictor
VQ  Vector Quantisation
VQEG  Video Quality Experts Group
WSNR  Weighted Signal-to-Noise Ratio

Units

bit/pixel  bits per pixel
bit/sec  bits per second
c/deg  cycles per degree of visual angle
c/sec  cycles per second
cd/m²  candela per square metre
dB  decibels
deg  degrees of visual angle
deg/sec  degrees of visual angle per second
Gbit/sec  giga-bits per second
Hz  hertz
kbit/sec  kilo-bits per second
min  minutes of visual angle
msec  milliseconds
Mbit/sec  mega-bits per second
sec  seconds
Publications

These are the publications on topics associated with the thesis, which have been produced by, or in conjunction with, the author during his Ph.D. candidacy:

Journal Articles


Conference Papers


Authorship

The work contained in this thesis has not been previously submitted for a degree or diploma at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signed: ................................................ Date: ..........................
Acknowledgements

As my journey through PhD-land comes to an end, there are many people to whom I would like to express sincere thanks. Firstly I would like to thank the two people who acted as my principal supervisors during my candidacy. Prof. Anthony Maeder, who was my principal supervisor during the first year of the PhD and my associate supervisor thereafter, provided the initial inspiration for my interest in image processing and vision models. His constant support and advice over the duration of my PhD was crucial to its successful completion. Dr. Neil Bergmann took over as my principal supervisor for the final three years of my candidacy, and he provided continual support, encouragement, and helpful advice, particularly when it was most needed during the last year. Prof. Kurt Kubik, as centre director and associate supervisor, provided inspiration and instilled confidence as he does in all his students, and I am particularly grateful for the financial travel support which he enabled. This support was continued by Dr. Mohammed Bennamoun when he took over as centre director. I would also like to express my gratitude to Dr. Donald McLean, for his thoughtful input, patience, and continual motivation.

Special thanks go to my fellow students at the Space Centre, for the friendship that they provided. In particular, I would like to thank Rob Reeves, Adriana Bodnarova, and Jasmine Banks, who “did time” with me in the Bare Geek Room. Though our technical discussions were important, my fondest memories are of the many humorous incidents which I was fortunate enough to witness: who could forget those massage demonstrations! I must give special thanks to Adriana for her thorough and diligent review of my thesis (which, in her opinion, reads like...
“a mystery”!, and also to Dr. Rod Walker for initially getting me interested in undertaken this PhD (I am grateful, now that it is over!). I am indebted also to the many secretaries and computer support people who have helped me during my time at QUT. I would also like to thank the School of Engineering at the University of Ballarat, for generously allowing me to use their subjective testing facilities.

I am appreciative of the financial support which I received during my candidacy, through the Australian Postgraduate Award and the CSIRO Telecommunications and Industrial Physics scholarships. The Space Centre also provided generous financial support during the final few months of my PhD, and I am particularly grateful to them for giving me the opportunity to travel to overseas conferences. The learning and exchange of ideas which ensued, and the important contacts which I made, were invaluable and would not have been possible without this support.

Doing an engineering thesis involves more than just algorithms, papers, and experiments: one must try to maintain sanity as a background process. For this reason, I am especially grateful to my family and friends. With my family providing continual love and encouragement, and my friends attending to the social aspects, I believe my sanity has remained intact! Special thanks to Dominique, for her love, patience, and understanding during my nocturnal stages, not to mention the gourmet lunchbox to which I was treated daily! Last but by no means least, I thank God for guiding me throughout this PhD. His presence was particularly evident during the hardest times, and He always provided strength when it was most needed.
Chapter 1

Introduction

During the last decade there has been a tremendous increase in the use of digital pictures. Some applications which have benefited from digital image processing techniques include medical imaging, remote sensing, factory automation, and video production. Digital representation of pictures has advantages over analog for several reasons, including:

- increased flexibility in the manipulation of the images,
- novel image processing applications are possible,
- higher quality pictures are achievable,
- digital data is more amenable for use by computers.

However, processing of digital images suffers from a major problem in the large number of bits required to represent the images. For instance, consider a standard colour HDTV signal displayed progressively at 60 frames per second, with a spatial resolution of $1280 \times 720$ pixels at 24 bit/pixel. At raw data rates, such a system would produce over 1.3 Gbit/sec. This rate is obviously unacceptable for
Introduction

transmission and storage for many applications. Thus the need for compression of pictures is apparent.

The number of applications requiring compression is continually increasing as more sophisticated use is made of the digital nature of the representation. Examples of specific applications which have been made possible through the use of compression include:

- videoconferencing and videophone,

- digital television,

- multimedia electronic mail,

- video on demand,

- digital picture archiving.

With the introduction of improved technology such as broadband optic fibre networks and large capacity storage devices, it may appear that compression could be unnecessary in the future. However there is significant evidence to suggest that compression will still play a major role in digital imaging systems in at least the short to medium term. The factors which support this are listed below.

- Although major network nodes will be linked via optic fibre, it is expected to still take a long time to connect all users via optic fibre due to the huge expense involved with connecting individual homes and offices.

- Compression will always be necessary for systems requiring wireless communication, due to inherent bandwidth limitations.

- The number of applications requiring compression is rapidly increasing, so increases in channel capacity can be expected to be followed by increases in the bandwidth required by these new applications.
Past trends indicate that although processing power and storage ability continuously improve at a tremendous rate, the increase in channel bandwidth occurs at a much slower rate, and typically in a more stepped rather than uniform manner.

Digital compression standards such as MPEG and JPEG have been accepted as industry standards for a variety of applications such as digital television and picture archiving. Consumer products utilising these standards and incorporating them as hardware are likely to be used well into the next century.

Daily lives are becoming far more dependent on data based services and activities, so the volume of data demanded by users is increasing substantially.

Thus it appears that there will continue to be an increasing need for compression in the foreseeable future.

1.1 Assessing the Quality of Compressed Pictures

The best lossless compression schemes (i.e. the reconstructed image is identical to the original) only offer compression ratios of between 1.5:1 and 3:1 for natural images. Although a few applications require lossless compression (e.g. some types of medical imaging), the majority of applications utilise lossy compression (i.e. the reconstructed pixel values may be changed), which offers significantly higher compression rates.

As soon as a compression scheme is used which introduces losses into the picture, it is necessary to be able to quantify the quality of the reconstructed picture.
1.1 Assessing the Quality of Compressed Pictures

Quality can take on many different meanings, depending on the application and the end user. For instance the quality of a scheme used to compress images in an automated face recognition application could be judged by the percentage of false detections introduced as a result of compression. However, in most applications a human viewer is the final judge of picture quality. In such cases it is advantageous to take into account the properties of the human visual system (HVS) when designing the image compression algorithm, and when assessing picture quality. As discussed in detail in Chapter 2, the limited processing capabilities of the HVS allow images to be significantly changed without the distortions being detectable or objectionable. This thesis focuses specifically on taking advantage of properties of the HVS affecting picture quality as judged by human viewers.

Digital image compression techniques introduce a wide range of different distortions including blockiness, ringing, blurring, and colour errors (see Section 3.1.4). The complex nature of these distortions means that quality is best defined as a multi-dimensional entity. A useful distinction which is commonly made is to separate the terms image fidelity and image quality. The fidelity of a compressed image defines the extent to which a difference can be detected between it and the original by human viewers. The quality of a compressed picture refers to the naturalness and the overall impression created by the picture, and extends beyond the mere detection of a visible difference. Quality judgments are strongly affected by the largest impairments in a scene, particularly if they occur in regions of interest (ROIs). Testing for image fidelity is far more simple, reliable, and systematic than testing for image quality. This is because image quality evaluation can require an assessment of the impact of several different and clearly visible (suprathreshold) distortions. Such assessments involve higher level perceptual and cognitive factors, which are often difficult to model and may vary depending on many factors such as observers, images, viewing conditions and viewing instructions.
Subjective testing (see Section 3.2) is currently used as a general approach to obtain the most accurate estimate of picture quality. However the practicality of such tests is limited since they are time consuming, expensive, difficult to repeat, and require specialised viewing conditions. These problems have resulted in the use of objective quality metrics (i.e. measures generated by algorithms which aim to correlate well with human viewer response). Simple objective metrics such as Mean Square Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) are commonly used due to their simplicity and to the absence of any standardised or widely accepted alternative. However it is well known that such simple metrics do not correlate well with human perception [44, 177, 257]. The development of more accurate and robust objective quality assessment techniques is currently an area of active research. However, at present there is no alternative technique which has been widely accepted by the image processing community.

Improved knowledge of the operation of the HVS has led to the development of objective quality metrics based on HVS models (e.g. [44, 153, 256]). The general approach is to model the early stages of the human visual system, from the optics through to the primary visual cortex. Model data is obtained from psychophysical experiments which determine detection thresholds for simple, artificial stimuli such as sine-wave gratings and Gabor patches. Care must be taken when applying these thresholds to natural images, since visual thresholds are known to vary significantly depending upon the stimulus and the test conditions (see Sections 4.2 and 4.3). These early vision models generally produce a map commonly referred to as a Just Noticeable Difference (JND) map. This map shows, for each point in the image, whether a difference between the original and coded picture is visible.

Human vision models show promise and have been successful in assessing the fidelity of images [44, 153]. This is particularly useful in high quality applications, where a visible difference occurring anywhere in the coded image is unacceptable. However, many compression applications allow suprathreshold levels of distortion
and thus require the assessment of quality rather than fidelity. In order to use HVS-based metrics to provide an accurate assessment of the quality of pictures coded with suprathreshold errors, other factors need to be considered. Current techniques for determining picture quality from JND maps generally involve a simple Minkowski summation of the visible errors, to produce a single number which is intended to correlate well with human opinion of overall picture quality. Such simple approaches fail to take into account higher level perceptual processes.

Visible errors of the same magnitude are treated as being of equal importance in these methods, regardless of their location in the image. This is in contrast to the operation of attentional mechanisms of the human visual system, which are known to allocate higher importance to visually salient regions in an image [135]. Our perception of overall picture quality is strongly influenced by the quality in the few ROIs in the scene [65, 96, 136, 252]. Other areas can undergo significant distortion without strongly affecting overall picture quality.

1.2 Scope of the Research

There is a strong need for accurate, robust, efficient, and objective means for assessing the subjective quality of compressed pictures. This thesis presents a new early vision model for monochrome images which improves upon previous models by specifically taking into account the structure of natural images and the compression process. This is important, since it has been shown that HVS models which are based on data using simple, artificial stimuli can perform poorly when applied to natural scenes [3, 74, 214]. This model is shown to provide an accurate indication of image fidelity, but is not powerful enough to capture all of the effects involved in picture quality evaluation. To extend this early vision model for use as a robust image quality metric, a model of human visual attention is used to identify ROIs within a scene. This is done through the use of Importance Maps (IM) [157, 190]. These maps identify areas in a scene which are strong
visual attractors, based on both bottom-up and top-down factors. These areas are likely to be the focus of attention in a scene, and the quality of the picture in these areas has the largest influence on overall picture quality [65, 96, 136, 252]. The IMs are used to weight the visible errors predicted by the early vision model. This essentially means that perceptible errors occurring in visually important parts of the scene will have a greater influence on overall picture quality than equally perceptible errors occurring in areas which are not the focus of attention. By taking into account the location of the errors, a more accurate indication of picture quality can be obtained. The results of correlation with subjective quality rating data are presented which confirm the significant improvement in prediction accuracy achieved by this model.

Perceptual vision models such as the one developed here are useful in many other areas of image processing. For instance, they can be used to provide adaptive control of the quantisation process in an image or video compression algorithm. This thesis demonstrates the usefulness of this approach by extending the IM model to video and using components of the early vision model to control the locally adaptive quantisation in an MPEG encoder. The results of subjective tests comparing the quality of sequences coded using this perceptually-tuned MPEG encoder with standard MPEG encoded sequences are shown, and confirm the improved subjective quality achieved using this approach. Further applications for visual models in image processing are identified and investigated.

1.3 Overview of the Thesis

The organisation of the thesis is as follows. Chapter 2 provides an overview of the properties of the HVS relevant to perceptual vision models which are designed for application with natural scenes. Both early visual processes (from the optics of the eye through to the primary visual cortex) and later visual processes (at-
tention, eye movements and cognitive factors) are discussed. The early stages of
Chapter 3 give a brief review of different compression techniques, outlining com-
mon compression artifacts. Later parts of this chapter review different subjective
and objective quality assessment techniques which have been developed. This is
followed by a critical review of objective quality metrics based on HVS models
in Chapter 4. This chapter also identifies problems with current HVS models,
 focusing on their application with compressed natural images.

In Chapter 5 a new early vision model for assessing image fidelity and quality
is presented. This metric is designed specifically for use with complex natural
images. The effects that individual model components have on the overall per-
formance of the early vision model are investigated. Chapter 6 details a method
for determining important regions in an image, by extracting several features
from the image which are known to affect visual attention and combining them
to produce an IM. Both a review of previous attention models and a detailed de-
scription of the IM technique are provided. These IMs are then combined with the
aforementioned early vision model to provide a novel image quality assessment
 technique in Chapter 7. The significant improvement in accuracy achieved by us-
ing this technique in comparison to other methods is demonstrated by correlation
with subjective quality rating data. Chapter 8 discusses ways of utilising percep-
tual vision models to control a compression algorithm. A perceptually-tuned
MPEG encoder is detailed as an example. This encoder requires an extension
of the IM model to include temporal information. In the discussion and conclu-
sion the merits of the vision models presented in this thesis are discussed, future
work suggested, and other image processing applications which could benefit from
perceptual vision models such as the one presented in this thesis are investigated.
Chapter 2

The Human Visual System

Most digital images are ultimately viewed and appraised by humans. It is therefore important for imaging systems to be designed with consideration of the operation of the HVS. This ensures that only information relevant for the required task is received by the viewer, thus avoiding unnecessary expense in image storage, transmission, and manipulation.

Knowledge of basic properties of the HVS has been used in a significant number of imaging applications in the past. For example, sub-sampling of chrominance components in analog television enabled a halving in transmission bandwidth, with minimal perceptual degradation. The improved understanding of the operation of the HVS which has been gained since the 1960s, and the increased interaction between engineers and vision scientists has resulted in more widespread usage of HVS properties in digital image processing systems. This trend appears likely to continue as improved visual models are developed, and as more applications requiring computational vision models are recognised.

Understanding the operation of the HVS is particularly important for the development of accurate and robust picture quality assessment techniques. Although alternative objective techniques have been proposed (see Section 3.3),
these techniques invariably suffer from inherent limitations which prevent their general applicability. Recently proposed metrics based on early vision models (e.g. [44, 152, 256]) offer no such inherent difficulties, and can accurately assess the fidelity of compressed pictures. Although they give a reasonable estimate of picture quality, further improvements can be made by extending these early vision models: firstly by tuning the early vision model specifically for complex natural images, and secondly by including some higher level visual processes. This is important because more complex cognitive processes such as attention are used when assessing the quality of a picture. This thesis examines and develops a visual attention model which is used in conjunction with an early vision model (tuned for complex natural images) to provide an improved objective image quality metric.

This chapter gives a review of the operation of the HVS. In particular, it focuses on properties relevant for vision modeling and picture quality assessment. Section 2.1 discusses the neurophysiology of the HVS, beginning with the eye and then following the pathways to the visual cortex and higher processing areas. In Section 2.2 some important properties of the early stages of the HVS (up to the primary visual cortex) are examined. The selective nature of our visual system is discussed in detail in Section 2.3, through the closely related processes of eye movements and visual attention. Both bottom-up (stimulus driven) and top-down (task driven) factors which are involved in the control of these processes are covered.

2.1 Optical and Neural Pathways

A cross-section of the human eye indicating the major components is shown in Figure 2.1. Light from an external object is focused by the cornea and the lens to form an image of the object on the retina. The elastic nature of the lens allows
focusing at different distances, which is called accommodation. Two types of photoreceptors are contained in the retina: cones, which are responsible for colour and acuity at normal daytime (photopic) light levels, and rods, which are used for low light level (scotopic) vision. Three different types of cones exist: L-, M-, and S-cones, which are most sensitive to long-, middle-, and short-wavelengths respectively. The presence of these receptors tuned to different wavelengths forms the basis of our colour vision. The rods and cones convert the incoming light into neural signals, and then transfer the signal through several synaptic layers and through to the ganglion cells and onto the optic fibre nerves. The retina is therefore considered part of the brain, the only part which can be observed in the intact individual. Each retina contains about 120 million rods and 6 million cones. However, the optic nerve contains only about 1 million fibres; thus significant processing occurs in the eye to reduce the bandwidth of information transferred to the brain.

In the centre of the retina is a small area called the fovea, which occupies about 2 deg. The centre of the fovea contains only densely packed cones and is thus the
Figure 2.2: Density of receptors in the retina as a function of eccentricity (adapted from [53, Figure 3.4]).

area of highest acuity. Figure 2.2 shows the distribution of cones and rods across the retina. This clearly demonstrates the high density of cones around the fovea, and the rapid decrease in cone density in peripheral regions. The density of rods is shown to be greatest at around 15 deg from the fovea, with gradual fall off with further eccentricity. Although the number of rods far exceeds the number of cones, later neural processing is concentrated on signals originating in the fovea, so cones provide the bulk of information to be processed by later brain areas. As an example, the receptor to ganglion cell ratio in the central fovea is around 1:3, while the average for the whole retina is 125:1. Note that the output of each ganglion cell is directly connected to an optic nerve fibre. Figure 2.2 also shows the blind spot where no rods or cones are located, which is due to the presence of the optic disk (the point where the optic nerve fibres attach to the retina).

The positioning of the fovea is changed by rapid, directed eye movements called
saccades, which occur every 100–500 milliseconds. The location of future saccades is controlled by visual attention processes, which search peripheral regions for important or uncertain areas on which to foveate. A full description of eye movements and attention processes is contained in Section 2.3.

The need for such a variable-resolution architecture in the HVS is apparent when one considers the amount of processing performed to obtain our high acuity foveal vision. Approximately 50% of the human visual cortex is exclusively visual [222], and a majority of this is devoted to the central few degrees of viewing. Schwartz [222] has noted that if our entire field of view possessed the same acuity as our fovea, we would require a brain weighing over 2000 kilograms. A variable-resolution architecture, which allows rapid changing of the point of foveation to peripheral regions of interest, therefore provides an efficient solution.

The pathway of the visual signal from the retina and through to the cortex is shown in Figure 2.3. At the optic chiasm, the nasal fibres cross over to the other side while the temporal fibres remain on the same side of the brain. This means that visual signals from areas to the left of the fixation point (left visual field) are projected onto the right lateral geniculate nucleus (LGN) (and the right primary visual cortex), while signals from areas to the right of the fixation point are projected onto the left LGN. Following the optic chiasm, the fibres are referred to as the optic tract. The destination of around 20% of the optic tract fibres is the superior colliculus [53, p. 68] (used in the control of eye movements), and a small number of fibres are used to control accommodation and pupillary movements. However, the majority of optic tract fibres go to the LGN.

At the LGN, the seemingly unordered bundle of fibres in the optic tract fan out in a topographically ordered way. The LGN contains 6 layers of cells which are stacked on top of each other (Figure 2.4). Three of the layers receive input from the left eye and three from the right eye. Areas in the same location in different layers receive input from the same location in the visual field; hence a
Figure 2.3: Visual pathway from the eyes to the primary visual cortex (adapted from [53, Figure 3.10]).

topographical map of the visual field is contained in the LGN layers. The bottom two layers contain larger cells and are referred to as magnocellular or M layers, while the upper four layers consist of smaller cells which are termed parvocellular or P layers. A significant functional difference between these layers is that the upper P layers are concerned with colour vision, while the lower M layers are not. Single cells in the LGN respond to light in a similar centre-surround way to ganglion cells (e.g. Figure 2.5(a)), so in terms of visual information there is no profound transformation occurring. However, the LGN does receive feedback from the cortex, which plays a role in attention or arousal processes [117, p. 62].

A majority of the visual signal from the LGN arrives at the primary visual cortex,
or area V1. This area is also referred to as the striate cortex, due to its layered or striped cross-sectional appearance. Although the LGN only contains around 1 million neurons, the primary visual cortex contains around 200 million neurons; hence significant computation and transformation of the visual signal occurs in this area. The primary visual cortex consists of crumpled 2 mm thick sheets of neurons, with a total surface area of around 1400 cm² [267, p. 153]. The primary visual cortex has been the subject of intense study, and a good understanding of the operation of its neurons has been gained [53, 117, 295].

Neurons at the first stage in area V1 respond with a circular symmetry (as in the retina or LGN), which means that a line or edge produces the same response.
Figure 2.5: Receptive fields of simple cells. (a) Centre-surround with varying sized excitatory areas, (b) orientation selective with odd symmetry, and (c) orientation selective with even symmetry. The centre-surround response in (a) is also found in ganglion and LGN cells.

regardless of its orientation (Figure 2.5(a)). However, by the second stage of area V1 (which receives its input from the first stage), the neurons are responsive to oriented edges and lines (Figure 2.5(b) and (c)). This characteristic was first discovered by Hubel and Wiesel in 1959 [118], when investigating the response of neurons in the cat’s visual cortex. Although most of the neurons in V1 exhibit this orientation sensitivity, their behavior differs in other ways.

A distinction that was made by Hubel and Wiesel was to classify the cells as either simple or complex. Simple cells have receptive fields similar to those shown in Figure 2.5(b) and (c). They show summation within the excitatory and inhibitory regions, such that a stimulus that covers all of the excitatory region produces a larger response than one covering only a part of the excitatory region. Simple cells also show an antagonism between excitatory and inhibitory areas, so that a light stimulus which covers both the excitatory and inhibitory regions produces a lower
response than if it were restricted to the excitatory region alone. Complex cells are more numerous than simple cells, accounting for around 75% of all cells in the striate cortex [117, p. 74]. It is believed that they receive their input from several simple cells, and their receptive fields are therefore different. Figure 2.6 shows a possible wiring diagram for a complex cell, and the resulting receptive field which is produced. Complex cells do not possess discrete excitatory and inhibitory sub-regions in their receptive fields: they give on-off responses to changes anywhere in their receptive field. However, they still respond best to bars which are a certain fraction of the total receptive field width, while giving no response to bars covering the whole receptive field. Complex cells respond strongest to moving stimuli, since they adapt quite rapidly to stationary stimuli.

As well as being sensitive to particular orientations, cortical cells are generally selective of particular spatial frequencies, directions of movement, velocities, number of cycles in the sine-wave grating stimulus, and lengths of gratings or bars. The sensitivity to these different stimulus features varies considerably across different cells. Orientation bandwidths of individual cells vary from around 6–8 deg to 90 deg (or 360 deg, if non-oriented cells are included) [53, p. 266]. The median orientation bandwidth of cortical cells is however around 35–40 deg. Cortical neu-
rons are sensitive to a wide range of spatial frequencies. The spatial frequency bandwidths of the neurons are also highly variable (from 0.5–3.2 octaves, with a median of around 1.0–1.4 octaves [53, p. 205]). Cortical cells generally respond best when a specific number of cycles are present in the sine-wave grating stimulus. Peak response varies from one to several cycles, and generally drops off if more than this number of cycles are present [53, pp. 111–112].

The anatomy of the striate cortex shows that it is organised into 6 layers, as shown in Figure 2.7. These layers vary in cell density and in their interconnection with the rest of the brain. Input from the LGN is received in layer 4. Magnocellular neurons send their outputs to 4Co, while parvocellular outputs are sent to 4Cβ and 4A. Layer 4B receives input from 4Co and sends its output to other cortical areas. Layers 2 and 3 receive input from 4Cβ and send their output to other cortical areas, as well as to layers 5 and 6. Layer 5 sends a major output to the superior colliculus, while layer 6 sends a large output back to the LGN [267, pp. 155–157].

The modular structure of the primary visual cortex is illustrated by the model in Figure 2.8. Although the physical cortex structure is not as regular as shown in this figure, the general modular nature can still be appreciated. Areas which are sensitive to particular spatial frequencies and orientations are arranged in orthogonal columns. These exist for both the left and right eye (ocular dominance columns). Each column has only a narrow spatial extent, but the receptive fields of cells within a column may overlap with those of neighbouring columns.

There are at least 20 visual areas other than the primary visual cortex. Much less is known about the operation of these areas, but they are believed to work in both a serial and parallel fashion. Some areas are known to be responsible for specific aspects of our vision. For instance, area MT (medial temporal) is sensitive to motion and stereo, while in area V4 the neurons are concerned with colour processing. However, serial projections through several visual areas are
Figure 2.7: Layered structure of the primary visual cortex, showing interconnections between the layers. The descriptions on the right of the figure indicate the predominant cell types found in each layer (adapted from [117, p. 99]).
Figure 2.8: Modular structure of the primary visual cortex. The orthogonal arrangement of cells tuned to specific spatial frequencies and orientations is evident (adapted from [53, Figure 4.23]).

also known to take place. In addition to these forward projections, considerable feedback also occurs from later visual areas to areas from which they receive input, and to other areas of the brain such as the superior colliculus and to various subdivisions of the thalamus (the thalamus is a complex group of mid-brain cells, of which the LGN forms a small part). Investigation into the functioning of these later cortical areas is currently an area of active research.

2.2 Properties of Early Vision

A remarkable feature of research into the HVS has been the convergence of results found from both neurophysiological and psychophysical studies. This section examines some of the most important properties of early human vision, giving both neurophysiological explanations for their existence and psychophysical re-
sults which quantify the effects of these properties. An underlying focus of this section (and of the thesis) is to determine, both qualitatively and quantitatively, the properties of human vision which are relevant and appropriate for the viewing of natural images. Thus some constraints on the types of stimuli which will be input to the HVS models are known and should be taken into account (e.g. luminance ranges of display devices; viewing conditions). Such constraints are reasonable and necessary, if one is to develop a useful, practical model of human vision.

The HVS is highly adaptive to the diverse and complex range of stimuli which it deals with in the natural world. For example, our cones are dynamically sensitive and linearly responsive over a range of only $10^2$–$10^3$. However, by adapting to the local luminance level, our cones can effectively operate over a much larger range. In combination with the scotopically sensitive rods, our sensitivity to luminance is increased to a range of $10^8$. This example demonstrates the adaptive nature of our visual system. It also suggests that over a limited range of inputs, visual processes can be sufficiently well approximated by (or transformed to become) linear functions. This linearity over a range is found in many aspects of vision, and greatly simplifies the modeling of visual processes. However, as will be seen later in this section and in Chapter 4, care must be taken in assuming linearity when applying psychophysical results from tests using simple, artificial stimuli to complex, natural images.

### 2.2.1 Sensitivity to Luminance and Contrast Changes

A common feature of the human visual system which could be seen in Section 2.1 is that it is not sensitive to the direct physical level of a stimulus, but to changes in the stimulus. Temporal changes are required since an image which is held stationary in the visual field for longer than 1–3 sec fades and disappears [291]. Eye movements and moving stimuli prevent this from occurring. Spatial changes
Figure 2.9: Typical luminance change required for detection of stimuli for different background luminance values (adapted from [99, Figure 13.11]).

are also necessary, since the visual system doesn’t code the exact luminance of a stimulus, but instead codes spatial changes of the stimulus with its background. This property can be explained physiologically by the centre-surround nature of ganglion cell receptive fields (see Figure 2.5(a)). A spatially invariant stimulus will evoke no response, since the net effect of the excitatory and inhibitory parts of the receptive field will cancel to produce a zero response. However, if there is some luminance change in the stimulus, the sum of excitation and inhibition will not be zero, so a response will be evoked.

An important characteristic of this coding of spatial change is that it is not the physical change in luminance that is encoded, but the ratio of the luminance of the stimulus with its background (termed contrast). This property is known as Weber’s law, and it holds over photopic luminance ranges. This effectively means that the visual system is less sensitive to a luminance change on a background of high luminance, than to the same change when the background is of low luminance. The relationship between the change in luminance needed for detection and the background luminance is shown in Figure 2.9. The Weber region can be seen to extend over photopic light levels. At lower background luminance levels, the threshold is proportional to the square root of the background lumi-
nance (DeVries-Rose region), while at even lower luminance levels there exists a linear region. However, natural images rarely contain areas in all three regions. For most natural scenes viewed in typical environments, only the Weber and DeVries-Rose regions need to be considered.

Since we are sensitive to contrast and not luminance, it is important to be able to define the contrast of different parts of an image. However, this task is more difficult than it may at first appear. Even for simple artificial images such as sine wave gratings or luminance disks, different definitions of contrast are used. A commonly used measure for the contrast of a single target on a uniform background is the Weber fraction:

$$C_{\text{weber}} = \frac{\Delta L}{L},$$ (2.1)

where $\Delta L$ is the increment or decrement in target luminance compared to the background luminance, and $L$ is the uniform background luminance. For periodic test patterns (e.g. the sine-wave grating and Gabor patch shown in Figure 2.10), the Michelson contrast is widely used:

$$C_{\text{Michelson}} = \frac{L_{\text{max}} - L_{\text{min}}}{L_{\text{max}} + L_{\text{min}}},$$ (2.2)

where $L_{\text{max}}$ and $L_{\text{min}}$ are the maximum and minimum luminance values respectively. These contrast definitions produce similar results for low contrast, simple stimuli. However, they suffer several problems when applied to more complex stimuli, as listed below.

- These two measures do not coincide (particularly at higher contrasts) and do not share a common range of values. Michelson contrasts range from 0 to 1.0, while Weber contrasts range from $-1.0$ to $\infty$.
- These measures define contrast for the whole scene, based on global scene properties. However, natural images have varying luminance (and hence
Figure 2.10: Typical stimuli used in psychophysical tests to measure contrast. (a) Full-field sine-wave grating, and (b) Gabor patch. Both of these stimuli have the same Michelson contrast.

contrast) over the whole image; hence a local definition of contrast is appropriate for natural images.

- These measures fail to take into account the spatial frequency content of the image. This is important, since human contrast sensitivity is highly dependent upon spatial frequency, particularly at the threshold of detection (see Section 2.2.2).

It is apparent that simple contrast measures such as Weber and Michelson are inappropriate for vision models dealing with complex natural images.

Peli [195] has defined a local measure of contrast which addresses these issues. His measure, termed local band-limited contrast (LBC), defines a value of contrast at each point in the image, and for each spatial frequency band. The LBC is given by:

\[ c_k(x, y) = \frac{a_k(x, y)}{l_k(x, y)}, \tag{2.3} \]

where \( c_k \) represents the local contrast in a particular octave-width frequency band \( k \), \( a_k \) is the band-pass filtered image centred on frequency band \( k \), and \( l_k \) is the...
local luminance mean (i.e. a low-pass filtered version of the image containing all energy below band $k$). Note that this original definition of LBC does not consider orientation selective channels, but these can easily be included if required (see Section 5.1.2).

Psychophysical tests have been performed using moderately complex sine-wave gratings and Gabor patches, to determine how closely various contrast metrics model human perception of contrast [197, 198]. Results from these tests showed that the LBC contrast metric, with a slight modification suggested by Lubin [153], was the only contrast metric from the group tested which agreed consistently with subjective data. The LBC metric even agreed with subjective data at suprathreshold contrasts, giving further indication of its potential for use in natural images.

### 2.2.2 Frequency Sensitivities

A convenient way to quantify the performance of the HVS is to perform psychophysical tests at the *threshold* of detection. This approach has physiological justifications, since neurons typically do not respond when the stimulus is below the detection threshold. A test which has been widely studied is to assess the contrast required for detection of a stimulus on a flat background. This value is referred to as the *contrast threshold*, and its inverse is called the *contrast sensitivity*. By using a sinusoidal grating as the stimulus, the threshold sensitivity at different spatial frequencies can be determined. This relationship is termed the *contrast sensitivity function* (CSF). Since any complex waveform can be analysed into a series of sine and cosine waves, the CSF provides a powerful way of determining the visibility of complex stimuli. However, this assumes that the visual system operates in a linear fashion. As will be discussed later in this section (and in Sections 4.2 and 4.3), various non-linearities exist in the HVS which limit the linear systems approach and require that it be applied with caution.
Typical spatial CSFs for a normal human observer at photopic light levels are shown in Figure 2.11, using data from van Nes [259] (full-field stimuli) and Rovamo et al. [155, 172, 215, 216] (patch stimuli). The dominant shape of the CSF can vary from low-pass to band-pass, and the peak sensitivity can also change significantly, depending on the nature of the stimulus and the viewing conditions. This is particularly apparent in Figure 2.11. As shown in this figure, the peak sensitivity when full-field stimuli were tested was around five times higher than when patch stimuli were tested, and the shape of the CSF was band-pass rather than low-pass. The strong high frequency attenuation found in all CSFs is due primarily to blurring in the optics. This includes factors such as cornea and lens imperfections, light scattering in the eye, diffraction by the pupil, and incorrect shape of the eyeball. In addition, the spatial sampling of the retinal receptors sets an upper limit to the detection of high frequency information. The low spatial frequency attenuation is predominantly due to the neural inhibitory interactions between centre and surround of retinal receptive fields [267, p. 135].
Figure 2.12: Variation of contrast sensitivity with arbitrary mean luminance values, $L_1 < L_2 < L_3 < L_4$ (adapted from [99, Figure 13.11]).

2.2.2.1 Factors which Influence the Spatial CSF

As demonstrated in Figure 2.11, both the shape and peak sensitivity of the CSF can change significantly depending upon experimental conditions. If the CSF is used as part of a HVS model for the viewing of natural images, it is important to choose a CSF which has been obtained using appropriate naturalistic stimuli and viewing conditions. The following list describes the experimental factors which have a strong effect on the CSF (see Sections 4.2 and 4.3 for a discussion of appropriate CSFs for natural image viewing).

- **Luminance.** Contrast sensitivity increases as mean luminance increases, although it reaches an asymptote at high luminance values [54]. The shape of the CSF also changes with varying luminance (see Figure 2.12). At low levels (typically $< 10 \text{ cd/m}^2$), the CSF is more low-pass in shape, with its point of greatest sensitivity at lower spatial frequencies of around $1$–$2 \text{ c/deg}$. This
is the DeVries-Rose region, where contrast threshold decreases as a power function with an exponent of 0.5. As the mean luminance level increases, the CSF for full-field stimuli becomes increasingly band-pass in shape, and the point of maximal sensitivity may increase to around 6 c/deg. The CSF is now operating in the Weber region, where contrast threshold stays constant. The transition between the DeVries-Rose and Weber regions is frequency dependent, with the mean luminance at transition being higher for higher spatial frequencies. Natural images typically contain both DeVries-Rose and Weber regions.

- **Spatial extent.** Increasing the spatial extent (either of a grating or an aperiodic stimulus) tends to increase the contrast sensitivity quickly, before reaching an asymptote at large spatial extents [215, 261]. Peli et al. [199] showed that peak sensitivity decreases by 0.5–1.0 log units for 1-octave width patch stimuli when compared to 4 deg × 4 deg fixed aperture gratings. A similar result was shown in Figure 2.11.

- **Orientation.** At low spatial frequencies, contrast sensitivity is virtually flat as a function of orientation. However, a small oblique effect exists at higher spatial frequencies (typically > 10 c/deg) [99]. This oblique effect results in sensitivity being reduced by a factor of 2–3 for gratings oriented at 45 deg, in comparison to vertical and horizontal orientations.

- **Retinal location.** Contrast sensitivity is typically greatest in the fovea, and decreases with distance from the fovea in an exponential manner [217]. Sensitivity decreases faster for higher spatial frequencies than for low. In general, the changes in sensitivity which occur with eccentricity are similar to those which occur with decreasing luminance level [53].

- **Temporal frequency.** The temporal frequency of a sinusoidal stimulus measures its rate of oscillation, in cycles per second. Spatial contrast sensitivity
Figure 2.13: Timing diagram which compares abrupt and gradual temporal presentation of stimuli.

is dependent upon the temporal frequency of the stimulus. This dependence leads to the concept of a spatiotemporal CSF. The spatial CSF is typically highest (and band-pass in shape) at medium temporal frequencies (3–10 c/sec), dropping off slowly (and becoming low-pass in shape) at low temporal frequencies, and dropping off quickly at high temporal frequencies [212]. The temporal frequency of the retinal image is dependent upon eye movements. Even when steadily fixating upon a point, the temporal velocity of the foveal image is approximately 0.1–0.2 deg/sec for unstabilised viewing [133]. The situation is further complicated when viewing natural images, since smooth pursuit eye movements (SPEMs) can track objects of considerable temporal velocity.

- **Temporal presentation.** Contrast sensitivity is also dependent upon the temporal aspects of the stimulus presentation (i.e. whether the stimuli were presented gradually or abruptly, as depicted in Figure 2.13). Peli [199] has shown that the CSFs for gradual onset stimuli are more band-pass in shape and have a slightly higher peak than for stimuli which are shown with an abrupt temporal onset.
• **Temporal Extent.** As the temporal extent (duration of stimulus presentation) is increased, contrast sensitivity dramatically increases at first, and then increases more slowly, until reaching an asymptotic value at temporal extents of a few hundred milliseconds [94].

Other factors which are known to have an effect of spatial contrast sensitivity, but which have less relevance for this work, include accommodation, colour, optics, and practice effects. A comprehensive review of factors which influence spatial contrast sensitivity can be found in [99].

### 2.2.2.2 Contrast at Suprathreshold

The shape of the CSF at threshold does not necessarily hold when contrasts are at suprathreshold levels. A contrast matching study was performed by Georgeson and Sullivan [95], which required subjects to match the contrasts of suprathreshold stimuli of different spatial frequencies. The results showed that as contrasts are increased, the perceived contrast function is virtually flat rather than band-pass across all spatial frequencies. This effect may be caused by post-striate processing. However, care must be exercised when comparing results from contrast matching and contrast threshold experiments, since they are inherently different procedures. Webster and Miyahara [274] demonstrated a strong reduction in the perceived contrast of low- and mid-frequency suprathreshold spatial gratings when the observers were adapted to naturalistic stimuli.

### 2.2.3 Masking

The previous section examined the ability of the HVS to detect stimuli on a uniform background. However when dealing with compressed natural images, both the stimulus (i.e. the coding error) and the background (i.e. the original
image) are typically non-uniform. It is therefore important to understand how the visibility thresholds of stimuli change when viewed on complex backgrounds. Stimuli viewed on a non-uniform background are generally less visible than when viewed on a uniform background. This reduction in visibility is termed masking: the presence of the non-uniform background masks the stimuli and increases the contrast required for their detection.

Spatial masking (commonly referred to as contrast masking) occurs when background luminance varies spatially. In a natural image, this is most obvious along an edge or in a textured region. Many psychophysical studies using either sinusoidal gratings or noise as maskers (e.g. [90, 140, 227]) have shown that the amount of masking which occurs is dependent upon the contrast of the background. An example of the types of stimuli used in these contrast masking tests is shown in Figure 2.14. The threshold of detection, which was determined from the CSF, is raised as the contrast of the masker increases. This effect is known as threshold elevation (TE), and its effect is demonstrated in Figure 2.15. The dashed line represents a facilitatory or “dipper” effect, which can occur near the threshold of detection if the stimulus and background are of similar spatial frequency, phase, and orientation [90, 173]. Masking typically begins when the contrast of the background equals the detection threshold. From this point, the threshold increases with an approximately constant gradient (on a log-log scale) of $\varepsilon$. The value of $\varepsilon$ can vary between 0.6 and 1.0, depending on the relation between the stimulus and background. This is discussed in greater detail in Section 2.2.3.3.

### 2.2.3.1 Spatial and Temporal Extents of Masking

Studies on the spatial extent of masking (i.e. the distance from a background discontinuity at which masking effects still occur) have generally shown that masking is a spatially localised effect, at least in foveal vision. Limb [143] examined mask-
Figure 2.14: Typical stimuli used in contrast masking tests. (a) Full-field sine-wave masker, (b) Gabor stimulus, and (c) superposition of masker and stimulus, which demonstrates the reduced visibility of the stimulus caused by the presence of the masker.

ing for foveal viewing at various distances from a luminance edge. He found that masking was greatest directly at the discontinuity, and slowly decreased until there was no masking at a distance of approximately 6 min from the edge. However for stimuli viewed peripherally, masking increased both in amplitude and in spatial extent (still present up to 20 min from the stimulus). This increase in the spatial extent of masking for peripherally viewed stimuli has also been demonstrated by Barghout-Stein et al. [15]. Polat and Sagi [203] tested the spatial extent of masking using Gabor patches as background and stimulus. A strong masking effect was found if the masker and the stimulus were separated spatially by up to 2 wavelengths, but the effect was not present if the separation
Figure 2.15: Threshold elevation caused by contrast masking.

was greater than this. Solomon and Watson [228], using Gabor patch stimuli and band-filtered noise as masker, obtained no masking if the masker-stimulus spatial separation was greater than 1 wavelength.

The temporal extent of spatial masking has been examined by performing detection experiments where the masker and the stimulus were not shown simultaneously; instead a variable length interval was placed between the displaying of the masker and the stimulus [89, 143]. Masking was strongest when the two stimuli were presented simultaneously, and dropped off quickly until the interval was 100–200 msec, when no significant masking occurred.

2.2.3.2 Spatial Frequency and Orientation Extents of Spatial Masking

It has generally been found that maximal masking occurs when the masker and the stimulus are of the same spatial frequency. However, there is some debate as to the extent of masking when spatial frequencies of the masker and the stimulus differ. Tests performed with full-field gratings of narrow bandwidth
(e.g. [52, 140, 279]) show that masking decreases as masker and test frequencies diverge, falling to zero when the difference is around ±2 octaves. However, when a background with wider bandwidth is used (i.e. band-limited noise), significant masking still occurs at ±2 octaves [150, 228]. Such wide-band masking may be more appropriate for models intended for complex natural images (see Sections 4.2 and 4.3). Vision models which test human detection performance in complex backgrounds [3, 74, 80] have obtained the highest accuracy in predicting human performance when a wide-band masking model was used. Another interesting feature of several masking studies [228, 243, 258] is the existence of a masking asymmetry. High spatial frequency backgrounds will mask low spatial frequency stimuli, but low spatial frequency backgrounds will not mask high spatial frequency stimuli.

Some disagreement also exists regarding the orientation extent of masking. The highest amount of masking has generally been found to occur when the orientation of the masker and the stimulus is the same (note however the facilitation effect shown in Figure 2.15). In tests with relatively long stimulus and masker durations (200–1000 msec), masking has been shown to occur only within a relatively narrow orientation tuning bandwidth of 15–30 deg [140, 202]. However recent data from Foley [90] using synchronous, brief (33 msec) stimuli and maskers shows that significant masking exists at all orientations of masker and stimulus. Such large variations in results once again emphasise the importance of using experimental data from tests that use stimuli which are appropriate for the application being tested.

2.2.3.3 Influence of Stimulus Uncertainty on Spatial Masking

The slope of the TE curve $\varepsilon$ (see Figure 2.15) depends not only on the contrast of the stimulus: it also depends on stimulus complexity and uncertainty. If an observer is uncertain as to the location, spatial frequency, orientation, shape,
size, onset, or duration of the stimulus, or if the background is cluttered and complex, then the task of detection is made more difficult, resulting in higher values of \( \varepsilon \). Masking experiments using sine-wave gratings as background and stimulus typically produce an \( \varepsilon \) of 0.6–0.7 (e.g. [90, 140, 202]). However when the background consists of band-limited noise, \( \varepsilon \) approaches 1.0 [150, 242].

Smith and Swift [227] showed that this significant change in \( \varepsilon \) is due to the increased uncertainty of the stimulus in the band-limited noise background. They demonstrated that, if subjects were repeatedly shown exactly the same band-limited noise pattern as a masker, then a learning effect ensued (i.e. uncertainty decreases) and \( \varepsilon \) would slowly drop from 1.0 down to a value of between 0.6–0.7. When naive observers were used for the sinusoidal masking test, they initially produced an \( \varepsilon \) value of around 1.0. However, the learning process was very quick for such simple stimuli, and within a few trials they became familiar with the sinusoidal stimuli (i.e. uncertainty was reduced) and the slope decreased to an \( \varepsilon \) value of around 0.6–0.7. A similar relation between the amount of masking and the learning effect has been found by Watson et al. [271], using a wider array of background types. The inclusion of uncertainty in a masking model for real world scenes has recently been advocated by Klein et al. [134].

He et al. [107] provides evidence that the complexity in a scene influences visual processing in areas past V1. Without distractors (i.e. low uncertainty), conventional visual resolution limits perception, but with distractors present (i.e. high uncertainty), perception of spatial detail depends on the ability of higher level attentional processes to isolate the items. Hence the influence that uncertainty has on our perception may be caused by visual attention rather than early vision processes.
2.2.3.4 Temporal Masking

It is well known that visual acuity is reduced if a stimulus is undergoing motion or change which is untracked by the eye [72, 87, 108] (i.e. retinal velocity > 0). This effect is called *temporal masking*. The amount of temporal masking induced by a stimulus is dependent upon its *retinal velocity* rather than its *image velocity*. Objects which are tracked by the eye show little if any temporal masking effects [72]. For a detailed discussion of eye movement characteristics, refer to Section 2.3.

Considerable temporal masking has also been found to occur around scene changes in image sequences. In most cases, resolution can be substantially reduced immediately following a scene change without the degradation being visible. However, the temporal extent of the masking effect is quite small (typically around 100 msec or 2–3 frames at conventional video rates) [233].

2.3 Eye Movements and Visual Attention

In order to efficiently deal with the masses of information present in the surrounding environment, the HVS operates using variable resolution. Although our field of view is around 200 deg horizontally and 135 deg vertically, we only possess a high degree of visual acuity in the fovea, which is approximately 2 deg in diameter. Thus in order to inspect the various objects in our environment accurately, eye movements are required. This is accomplished by saccades which rapidly shift the eye’s focus of attention every 100–500 msec. The way in which saccades are directed is by no means random, but rather they are controlled by visual attention mechanisms. Our visual attention is influenced by both *top-down* (i.e. task or context driven) and *bottom-up* (i.e. stimulus driven) factors. Attentive processes operate in parallel, looking in the periphery for important areas and uncertain
areas for the eye to foveate on at the next saccade. This section examines eye movements and visual attention processes, and demonstrates the strong relation that exists between them. Factors which are known to influence eye movements and attention are also discussed in detail.

2.3.1 Eye Movement Characteristics

Although eye movements are necessary for bringing areas of interest into foveal view (and this work is concerned primarily with these types of eye movements), they are also essential at a more fundamental level. While we may not be consciously aware of it, our eyes are undergoing constant motion. Even when we steadily fixate upon an object, small involuntary eye movements continue and ensure that the visual image never stabilises on the retina. This is essential since any image which is stabilised on the retina will fade and eventually disappear within 1–3 sec [291]. This occurs due to lateral inhibition effects in the retina, the purpose of which are to reduce the amount of information transmitted by only responding to changes in an image (both spatial and temporal). Kelly [133] has shown that a fixated object maintains a retinal velocity equivalent to 0.1–0.2 deg/sec.

A number of different types of eye movement have been identified, and their purposes and characteristics are discussed below. Some of these eye movements can be recognised in the eye movement recording shown in Figure 2.16.

- **Tremors.** These are extremely small amplitude (1.0–1.5 cone diameter), high frequency (70–90 Hz) oscillations, which are believed to be of minimal visual significance [42].

- **Drift.** This is characterised by slow, continuous movement of the eye which causes the image of the fixated object to drift away from the centre of the
Figure 2.16: Eye movements for a subject viewing a natural image. (a) Horizontal, and (b) vertical movement. Saccades and drifts are clearly visible.
fovea. The main purpose of drift is to prevent stabilisation of the image on the retina [291].

- **Vergence.** In vergence eye movements, the two eyes move in opposite directions in order to fixate on objects which are closer or further away from the observer [194].

- **Saccades.** Saccadic eye movements are rapid movements which typically occur 2–3 times every second, and may be instigated either voluntarily or involuntarily. Their purpose may be either to direct our point of foveation to a new ROI, or to re-foveate on the current ROI if the drift has become too large. The amplitude of the saccade can therefore vary considerably, but it rarely exceeds 15–20 deg [291]. Saccade duration will also vary with amplitude, and is typically in the range of 10–80 msec. During a saccade, extremely high velocities of up to 1000 deg/sec may be reached [109]. This high velocity would obviously cause severe blurring of the retinal image. However this blurring is not perceived due to *saccadic suppression*, a neural process whereby acuity is reduced during a saccade [267, p. 374].

- **Smooth Pursuit Eye Movements (SPEMs).** These occur when a moving object is being tracked by a person. Aside from the obvious difference in velocity, there are many other factors which distinguish SPEMs from saccades. Acuity is not suppressed during SPEM, as long as the object can be accurately tracked [72, 87]. SPEMs can only be generated in the presence of an appropriate stimulus; without such a stimulus a panning of the eyes results in a series of short saccadic movements rather than smooth pursuit movement [194]. Under ideal experimental conditions, observers can smoothly track objects with a velocity of over 100 deg/sec [291]. However when the motion is unpredictable or if the object is accelerating, the accuracy of SPEMs drops significantly [72]. When a new object appears in the visual field, there is typically a delay of 0.1–0.2 sec before SPEMs
begin to track the object [291]. Natural scenes typically contain several objects moving in a complex manner, and are displayed on a monitor with limited spatial extent. The predictability and trackability of objects in natural scenes is therefore significantly lower than that achieved under optimal experimental conditions.

### 2.3.1.1 Eye Movements while Viewing Natural Scenes

The development of eye tracking devices has enabled the analysis of subjects’ eye movements while viewing natural scenes. This has been the focus of many studies (e.g. [10, 31, 40, 149, 156, 182, 231, 280, 291]). A number of important characteristics of eye movements have been identified, and are discussed below.

- Fixations are not distributed evenly over the scene. A few regions typically attract the majority of fixations. These regions are generally highly correlated amongst different subjects, as long as the subjects view the image in the same context and with the same motivation. This has been shown to be true both for still images [31, 156, 291] and video [231, 280]. Mackworth and Morandi [156] found that two-thirds of fixations occurred in an area which occupied only 10% of the image area (20 subjects tested), and Buswell [31] has shown similar skewness of fixation distribution across a wide range of scenes (55) using 200 subjects. Even stronger correlation has been found for video in a study by Stelmach et al. [231]. Their study of the eye movements of 24 subjects for 15 forty-five second clips of typical NTSC video showed that well over 90% of fixations occurred in one of three ROIs in any scene, where an ROI occupied no more than 6% of total screen area. In many scenes (typically those with high motion), fixations were concentrated on just one or two ROIs. This suggests that eye movements are not idiosyncratic, and that a strong relationship exists between the direction of
gaze of different subjects, when viewing an image with the same context and motivation.

- Even if given unlimited viewing time, subjects tend not to scan all areas of a scene, but tend to return to previously viewed ROIs in the scene [182, 291].

- Subjects tend to be attracted to informative, unusual, or complex areas in a scene [10, 19, 149, 156]. In general, regions which stand out from their surrounds with respect to attentional factors are likely to be fixated upon. A detailed discussion of both bottom-up and top-down factors which attract our attention is contained in Section 2.3.4.

- The order in which objects are viewed in a scene may be repeatable by a subject, but not between subjects [182, 291]. The scanpath theory of Noton and Stark [182] gives evidence that individual subjects may regularly repeat the order of points fixated for a particular scene. However, the patterns of viewing of different subjects typically show limited correlation, even though the overall areas of fixation are highly correlated amongst subjects.

2.3.2 Visual Attention

The real world presents an overwhelming wealth of information to our senses, which is far beyond the processing capacity of our brain. In addition, short-term memory limitations generally allow us to store only $7 \pm 2$ ‘chunks’ of associated information at any time [169]. Attention provides us with a mechanism for performing some sort of selection, in order to focus on the objects or tasks which are most important or salient. In this sense, attention can be seen as a process which provides an additional means of information reduction, on top of that occurring in the variable-resolution retina and ganglion processing. William James, the 19th century psychologist regarded as the father of American psychology, provided the following definition of attention [124, pp. 381–382]:

41
“Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatterbrain state . . .”

Our visual attention is believed to contain both pre-attentive and attentive (focal) processes. Pre-attentive processing implies a spatially parallel system of very high capacity, which searches in peripheral regions for important or uncertain areas. Attentive processing is then used to provide concentrated processing resources to the location of interest. This attentive processing generally operates in a serial manner, although evidence does exist of divided attention (see [194, pp. 101–165] for a review of divided attention). Posner [207] has likened attention to a spotlight, which enhances event detection efficiency within its beam. Areas outside the spotlight of attention are processed in a very limited manner. This has been confirmed by memory tests, which show that subjects have little recall of stimuli which were not being attended [194]. The spatial extent of this spotlight has been shown to be variable; however the overall bandwidth of processing remains constant. Therefore, attention will not be as strong if it is spread over a large spatial area. Julesz [128] has shown that the location which the attention spotlight focuses on changes every 30–60 msec, meaning that attentional shifts occur several times faster than eye movements. However, another study by Duncan et al. [68] suggests that attention shifts occur much more slowly (up to several hundred milliseconds) for randomly blinking stimuli.

Neurophysiological studies have shown that a large number of brain areas are involved in visual attention [57]. Lesion and imaging studies reveal that the thalamus (and the pulvinar in particular) is involved in the engaging of attention, the
superior colliculus is responsible for moving attention, while the parietal cortex is active in disengaging attention. There is a significant increase in blood flow in attended areas in the early visual cortex, which emphasises the important role that attention plays even at the early stages of vision.

Like saccadic eye movements, attention can also be controlled in either a voluntary or an involuntary manner. Involuntary (transient) attention capture is generally quite fast, and is typically driven by bottom-up (stimulus driven) factors. Yantis [290] further categorises these types of attention capture as being either “strongly involuntary” (will still draw attention even if observer tries to ignore it) and “weakly involuntary” (can be ignored by the observer if they want to). Stimulus uncertainty has been shown to have a strong influence on involuntary attention capture. Visual search tasks using stimuli with spatial uncertainty are much more likely to produce strong involuntary responses than tests using stimuli where the observer is certain about spatial location (see [180]). This is significant in natural scenes, where stimulus uncertainty is often high. Voluntary attention capture on the other hand is generally much slower, and is usually driven by top-down influences.

Another feature of our attention system is that after a location has been attended, there is a tendency for the visual system to avoid shifting its focus of attention back to that area for a short period of time [206]. This is termed Inhibition of Return (IOR), and has a temporal duration ranging from 0.3–1.5 sec.

2.3.2.1 Methodologies Used to Research Visual Attention

Visual attention has been recognised as an important area of research by psychologists, since it is one of the only perceptual systems that directly involves both high level effects (such as memory and motivation) and low level processes. Visual search experiments have been widely adopted for the investigation of attention processes. In visual search tests, subjects search for a target item among
Figure 2.17: Sample stimuli used in visual search experiments. (a) Target (i.e. small dot) present, and (b) target absent. This particular test could be used to measure the effect of object size on visual search. The set size is 15 for both stimuli.

a number of distractor items. An example of the kinds of stimuli used in these tests is shown in Figure 2.17. The set size refers to the total number of items presented in the display, including repeated items. In half of the tests the target is present and in half the target is absent, and subjects must indicate whether they found the target or not. Two measures are commonly used to assess performance. The first is reaction time (RT). This can be plotted as a function of set size for both target present and target absent cases, and the mechanisms of the search can be determined from the function’s slopes and intercepts. Tests where the RT increases significantly with set size are termed inefficient, while tests showing RTs which are independent of set size are regarded as efficient (see Figure 2.18). The second measure is accuracy, which is a measure of how often the subject’s response is correct. In these experiments, the stimulus is presented for a varying length of time, and the accuracy of detection is plotted as a function of presentation time.

Extensive visual search tests have been performed for the purpose of identifying basic features. These are defined as features which support both efficient search
including visual search. In complex scenes, where a number of features may be influential,
reducing the influence of non-homogeneous surroundings, clutter, and
efficient search is significantly reduced if there is
unknown. However, the efficiency of search is significantly reduced if there is
knowledge of the stimuli or when the stimuli
with respect to a basic feature are able to attract our attention. This is the
moderate of attention processes, since stimuli which differ from their surroundings
and effortless segmentation [32]. Basic features are particularly important for the
and effortless segmentation [32]. Basic features are particularly important for the

Figure 2.18: Search slopes for efficient and inefficient search.
2.3.2.2 Modern Theories of Visual Attention

In his review of recent theories of attention, Allport [8] suggests that one of the main problems with current theories is that they try to propose a unitary theory of attention to explain all aspects of attention. Allport postulates that such an idealistic approach is unreasonable, due to the many different kinds of attentional functions which exist.

Despite their inability to explain all aspects of attention, a number of theories have been proposed which can explain a significant amount of visual search data. One of the best known of these is Treisman’s feature-integration theory (FIT) [249]. This theory suggests that visual search operates in parallel when a target differs from its distractors by a single feature (e.g., red target amongst green distractors), but must operate serially whenever conjunctions of two or more features characterise the objects (e.g., red ‘X’ amongst red ‘T’ s and green ‘X’ s). This theory therefore suggests a serial/parallel dichotomy of attention. The FIT has more recently been revised [248] to take into account new evidence which shows that efficient conjunctive search can occur under particular conditions [175].

More recent evidence casts doubt over the existence of a strict serial/parallel dichotomy of search. The original FIT suggests that individual features are searched in parallel. However, if the difference between the target and distractors is decreased, the search time increases in a smooth manner and eventually becomes serial [284]. This serial search stage occurs when the target is still clearly detectable (e.g., [174]), meaning that the JND for efficient search (termed a pre-attentive JND) is larger than the JND for detection. The serial search times found for conjunction searches also appear to be an artifact of the testing conditions used. A number of studies have found a continuum of search times for
conjunction searches (see Wolfe [284] for a review). Most recent theories of visual search therefore avoid the serial/parallel dichotomy but instead employ a continuum between the two extremities of serial and parallel search. Search becomes easier as the difference between target and distractor increases. There is however evidence of a saturation point, such that if the difference is increased beyond that point, no further increase in efficiency occurs [174].

The continuum of search has been the basis of the attention theory of Duncan and Humphreys [67], and also of the Guided Search theory of Wolfe [281, 285]. In Guided Search, bottom-up feature maps are generated for a number of basic features (e.g. colour, orientation, contrast). These features are under the control of top-down influence, and are combined to produce an overall activation map which indicates where attention is likely to be focused in the image for the specific task. The Guided Search theory has much in common with the selective attention models of Koch and his colleagues [135, 178, 180]. Their models combine a number of low level feature maps to produce an overall saliency map which indicates the salience of the different parts of the scene in terms of several bottom-up factors. A review of computational attention models is contained in Section 6.1.

### 2.3.3 Relationship Between Eye Movements and Attention

Our eye movements and attention are regulated by different, though closely interrelated, control systems. As mentioned in Section 2.3.2, changes in the location on which the attentional spotlight focuses typically occur much more frequently than saccadic eye movements. At some time during a fixation, visual attention is reduced in foveal regions and is allocated to peripheral regions to determine the location of the next fixation. Various top-down and bottom-up factors (discussed in detail in Section 2.3.4) are involved in determining the location of the next
fixation.

It is well known that we can attend to a peripheral object without actually making an eye movement to the object. This is referred to as covert attention, whereas changes in attention which involve saccades are termed overt. Posner [205] has shown that search tasks which require only limited acuity can successfully be performed using covert attention shifts. However, when attention is not foveally centred, acuity and performance diminish appreciably. Real-world viewing typically demands high acuity, and therefore requires overt shifts of attention.

Although shifts in attention can occur without producing any eye movements, the opposite is not true. Eye movements to a particular location are always preceded by a shift in visual attention to that location, and this coupling of eye movements with attention is mandatory [113, 194, 224]. This phenomenon is called the mandatory shift hypothesis, and occurs for both saccadic and smooth pursuit eye movements.

### 2.3.4 Factors which Influence Attention

Visual search experiments, eye movement studies, and other psychophysical and psychological tests have resulted in the identification of a number of factors which influence visual attention and eye movements. In the visual search paradigm, these factors are typically classified as being either top-down (task/motivation driven) or bottom-up (stimulus driven), although for some factors the distinction may not be so clear. In a sense, the simple bottom-up features may provide stronger attraction since they are capable of providing efficient selection, whereas higher-level or derived features generally do not (see [56, 241] for a discussion). It is difficult to search for targets which are not the most visually salient (in terms of bottom-up features) in a scene [22, 26, 248]. In addition, it is impossible not to attend to highly salient parts of a scene during a search task, if there is some
uncertainty regarding the location of the stimulus, even if the subject is given top-down instruction (see [180, 240]). Therefore, top-down instruction cannot override the influence of strong, bottom-up salient objects.

Another important finding regarding visual attractors is that we are attracted to objects rather than locations. Considerable evidence exists that our pre-attentive vision performs a segmentation of the scene into objects (see [247, 272, 284] for a discussion). Wolfe [282] has found that subjects have difficulty searching for targets where a segmentation of the scene was difficult, but that search efficiency improved considerably when a segmentation of objects in the scene was possible. Hillstrom and Yantis [111] have shown that it is the appearance of a new object rather than motion per se which captures our attention, and that it must be possible to segment the object pre-attentively. Zelinsky’s eye movement experiment [296] suggests a similar preference for objects. During his test, subjects’ saccades were directed at objects in the scene rather than particular locations in the display.

The remainder of this section discusses the various features which are known to influence visual attention individually, and concludes with an examination of the inter-relation of these factors.

2.3.4.1 Motion

Motion has been found to be one of the strongest bottom-up influences on visual attention [59, 175]. Our peripheral vision is highly tuned to detecting changes in motion, and our attention is involuntarily drawn to peripheral areas undergoing motion which is distinct from its surrounds (in terms of either speed or direction). The influence of motion has been found to be strongly asymmetric: it is much easier to find a moving stimulus amongst stationary distractors than it is to find a stationary target amongst moving distractors [59]. It is also easier to find a fast
target amongst slow distractors than vice versa [122]. There is evidence that the appearance of new or flashing stimuli may attract our attention rather than the actual motion per se [85, 111, 186]. In natural scenes, the predictability of the motion should also be considered, since unpredictable or accelerating motion is difficult to track and therefore produces a significant retinal velocity [72].

2.3.4.2 Contrast

The contrast of a region with its local surrounds has been found to exert a strong influence on visual attention. This has been found to be true in experiments using artificial [14, 26, 85, 126] and naturalistic [39, 229] stimuli. Efficiency of search increases monotonically with increasing contrast, but it appears that at very high contrasts a saturation point is reached, beyond which further increases in contrast do not improve search efficiency [126, 174]. Absolute luminance level has generally been found to have a small or insignificant effect on attention [39, 186, 291], which makes sense considering that luminance is converted to contrast at an early stage of visual processing. Examples of typical stimuli used in visual search experiments investigating the effect of contrast are shown in Figure 2.19.

2.3.4.3 Colour

Colour is another bottom-up feature which has been found to induce efficient visual search [14, 18, 38, 69, 114, 119, 174]. In many ways, the influence of colour is similar to that of contrast. Colours which are different from their surrounds will attract attention (and provide efficient search) once the difference from the surrounds increases above a certain level. As with contrast, there appears to be a saturation level, beyond which further increase in the colour difference provides no increase in search efficiency [174]. In cases where there are a number of distracting colours, efficiency can still be high if the target colour is sufficiently dissimilar.
Figure 2.19: Sample stimuli used to determine the influence of contrast on visual search. The target in (a) has a lower contrast than the target in (b), if contrast is measured with respect to the white background.

from the distractors. There is also evidence to suggest that particular colours (e.g. red) will attract our attention, or induce higher amount of masking than other colours [114].

2.3.4.4 Size

The size of an object has been shown to affect visual attention. Objects which have a different size from surrounding objects attract our attention [22, 23, 85, 119, 126, 176]. This relationship is quite strongly asymmetric, since larger objects amongst a background of small distractors will stand out much more than a small object amongst larger ones [22, 126, 176] (although Cole [39, 40] showed a smaller asymmetrical effect than other experiments). This asymmetric bias towards larger objects is demonstrated in Figure 2.20. Navon’s study [176] using stimuli with both global and local structure indicates that larger, global objects are attended to and processed earlier than smaller, local objects. Spatial frequency is also closely related to size and allows effortless segmentation and efficient search, but the influence of high spatial frequencies reduces significantly in the periphery [85].
Figure 2.20: Sample stimuli used to demonstrate size asymmetry in visual search. (a) Small target amongst large distractors, and (b) large target amongst small distractors. Search is typically more efficient in (b) than in (a).

2.3.4.5 Eccentricity and Location

Both visual search and eye movement experiments have indicated that the location within the viewing area has a strong influence on visual attention. Visual search experiments indicate that salience decreases with eccentricity, and that search targets are found quicker and more accurately when near the point of fixation [33, 85, 114]. Search for objects in real-world scenes also indicates a strong preference for centrally located objects when conspicuity and memory recall were tested [22, 39]. This central preference has been supported by several eye movement studies, which show that more [25, 40, 78] and longer [78] fixations occur in the central part of the scene. A strong preference for central locations has also been found when studying eye movement patterns during natural image [31] and video [75, 280] viewing, across a wide range of different scenes. Some of this effect is due to the fact that the camera is generally positioned to place the ROIs near the centre in the screen. However even in cases where there is no distinct ROI in the scene (e.g., repetitive patterns, maps), there is still a significant bias toward the centre of the image.
2.3.4.6 Shape

A factor which is difficult to study quantitatively is shape. This is probably due to the fact that shape is not easily defined, and contains many attributes. Thus many aspects of shape may require additional higher level processing, and may not be available pre-attentively. Some basic shape attributes have however been studied and their influence on visual attention has been determined. Line ends appear to be a basic feature, and shapes with line ends are generally more efficiently searched for than those without line ends [248]. This may be linked to evidence suggesting that closure may also be a basic feature [76, 77]. These experiments showed that objects with closed shape can be efficiently searched amongst open-shaped distractors. There is also evidence to suggest that corners and angles attract our attention. An eye movement study on natural images by Mackworth and Morandi [156] showed that subjects tended to fixate on angles and corners rather than simple or predictable contours. This agrees with the assumption that eye movements aim to minimise uncertainty in an image, so only irregular or unpredictable areas need fixations. Other important features of shape which influence attention are orientation and curvature (see [248, 284] for a review). Lines which differ from their surrounds with respect to orientation or curvature attract our attention and enable efficient search. In summary, shapes which are unusual, unpredictable and contain line ends and corners are likely to attract our attention. In natural images, long and thin objects, and objects with unusual shape, are therefore more likely to attract attention than objects with homogeneous, smooth, or predictable shape.

2.3.4.7 Foreground/Background

The 3-D depth of objects in a scene has a strong influence on attention. Depth can be determined either through pictorial cues (e.g. occlusion, relative motion, texture perspective) or through stereoscopic information. In both of these cases,
depth is a feature which allows a foreground object to be efficiently searched for amongst background distractors (see Wolfe [284] for a review). Various studies which examine both conspicuity and viewer eye movements during the viewing of natural images indicate a strong bias towards foreground objects in the scene [23, 31, 40].

2.3.4.8 Edges and Texture

A number of studies indicate that high contrast edges are strong visual attractors, particularly if their structure is unpredictable [32, 58, 156, 291]. However there is some debate regarding the influence of texture on visual attention. Although visual search experiments indicate that different textures can be searched for efficiently, eye movement studies for natural images suggest that textured regions do not necessarily receive a high number of fixations [156, 291]. Rough (unpredictable) textures however appear to be fixated on more than smooth textures.

2.3.4.9 Prior Instructions and Context of Viewing

The instructions or motivation that a subject has prior to viewing a scene (i.e. top-down factors) are known to have a very strong influence on visual attention and eye movements [225, 229, 291]. For example, a subject who is asked to count the number of people in a photograph would have significantly different eye movements compared to a subject who is asked to remember the clothing that the people in the photograph are wearing. In most visual search experiments, subjects are given an indication of the nature of the target which they are searching for. This produces a significant decrease in search time, since uncertainty about the target is small [39]. If no knowledge of the target is known a priori, then attention is driven primarily by the bottom-up salience of the targets. Significantly, even in cases where a visual target in known a priori, an object with strong low-level
visual salience will force attention to be allocated to it (see discussion of top-down and bottom-up factors in Section 2.3.4).

2.3.4.10 People

Another strong influence on visual attention in a scene has been found to be caused by the presence of people. Faces and hands in particular attract significant attention [223, 265, 291]. From a very early age, humans orient their attention to faces [127]. A study of eye movement patterns while looking at faces by Walker-Smith et al. [265] shows that people are strongly attracted to the eyes, nose, and mouth (although individual idiosyncratic patterns of facial viewing do exist).

2.3.4.11 Gestalt Properties

Many of the Gestalt laws of perceptual grouping also appear to play a strong role in influencing visual attention (see Bruce and Green [27, pp. 97–128] for a review of Gestalt laws). Some of these have already been mentioned (closure, orientation, proximity). Similarity may also have an influence on attention. For example, we may look at all four corners of a rectangle in order, or we may look for objects of a particular type in sequence, depending on top-down influence. Symmetry may also have an influence on attention. If an object is symmetric, then only one side will require attention: the other side can be inferred from its neighbour. Adaptation also appears to influence attention. Subjects will become less sensitive to a stimulus which is spatially or temporally repetitive, since the uncertainty of the stimulus is low.

2.3.4.12 Clutter and Complexity

Simple visual search tests using artificial stimuli have generally been performed on homogeneous backgrounds. This is unlike the situation found in typical natural
scenes, which often contain complex, cluttered, or heterogeneous backgrounds. Visual search tasks in more complex environments generally indicate that search is considerably more difficult [18, 22, 39, 107, 114, 119, 225, 282]. However the extent to which search is made more difficult is strongly influenced by the similarity and proximity of the target and its distractors. A complex background may not significantly affect search and attention if the target is dissimilar from its distractors, and also is highly salient. Wolfe [282] suggests that targets which can be pre-attentively segmented from their surrounds produce efficient search, while targets which cannot be easily segmented result in inefficient search.

2.3.4.13 Unusual and Unpredictable Stimuli

Another common theme in visual search and eye movement experiments is that unusual or unpredictable stimuli are likely to attract attention [10, 19, 149, 156, 204, 291]. This feature has been given many names by these different studies, such as: unusual, unpredictable, uncertain, informative, high information content, and different from surrounds.

2.3.4.14 Interrelation Between Basic Features

Although many factors which influence visual attention have been identified, little quantitative data exists regarding the exact weighting of the different factors and their inter-relationship. Some factors are clearly of very high importance (e.g. motion), but it is difficult to determine exactly how much more important one factor is than another. A particular factor may be more important than another factor in one image, while in another image the opposite may be true. Due to this lack of information, it is necessary to consider a large number of factors when modeling visual attention [161, 180, 250, 299]. It is therefore important for the attention model to identify areas of the scene which stand out strongly with
respect to a particular factor, and allocate high salience to that area. However if a particular factor has quite uniform distribution across the scene, then the significance of this factor in determining overall salience in that scene should not be very high. In this way, an automatic scene-dependent weighting of the features can be developed.

A few studies have looked at summation effects between different features. The effect of redundantly adding a new feature to a target so that the target is unique with respect to two features has produced mixed results. If the target is already strongly salient with regards to a single factor, then redundantly adding a second factor has limited effect on attention. However, if the target and distractors are quite similar, then introducing a second factor typically produces summation between the factors \([126, 174, 186]\). This result is based on limited data, so generalisation across all factors is dangerous. For example, Nagy and Sanchez \([174]\) found strong summation between luminance and colour when the target was darker than the distractors, but failed to find summation when the target was brighter than the distractors, except for when the target was blue in colour. Such interesting findings highlight the complex relationships that exist between individual factors. Visual search using stimuli made up of a conjunction of basic features also provides a range of efficiency results (see \([284]\) for a review). This effect however will not be investigated in this thesis, since stimuli containing such conjunctions are rarely found in the real world \([246]\).
Chapter 3

Assessing the Quality of Compressed Pictures

Virtually all consumer-oriented image and video applications rely on lossy compression as the mechanism for reducing the number of bits used by pictures to practically manageable levels. The worldwide acceptance of compression standards such as JPEG, MPEG, and H.261 has provided a means for conformance and interoperability of compression techniques. This is essential, since the proprietary compression techniques previously used were hindering the growth of consumer-based imaging products. The abovementioned standards are all based on the discrete cosine transform (DCT) and therefore produce similar types of artifacts. However, compression algorithms are still an area of active research. A variety of techniques (with different artifacts) are likely to be included as parts of future standards such as MPEG-4 and JPEG-2000.

The primary task of most compression schemes is to produce the lowest bit-rate picture (subject to various constraints) which still meets the quality requirements of the application for which it is designed. A means of quantifying the quality of the compressed picture is therefore required. In this thesis, relevant applications
are those in which a human viewer is assumed to be the final judge of picture quality.

Quality assessment of pictures distorted by analog transmission and conversion processes can be effectively performed using static test signals. However, such simple approaches are not suitable for digitally compressed pictures due to the dynamic and complex way in which distortions are introduced by various algorithms [300]. New techniques are therefore required for the assessment of digital picture quality.

A good picture quality measure has the potential to be used for many purposes, which are outlined below.

- It allows a coding algorithm to be tested against others at various compression rates, in order to rate the coder’s performance in comparison with competing coders.
- It shows the designer of the coding algorithm the exact location of major distortions in the picture, thus providing important feedback in the design process.
- It enables the coding algorithm to be analysed using a variety of test scenes and viewing conditions, so that its real-world performance may be assessed.
- It may be used to show that the coding technique passes certain performance criteria.
- It may be incorporated into the design of the coding algorithm (using either a feed-forward or feed-back approach), so that the perceptual distortion can be minimised. See Chapter 8 for a review of adaptive coders optimised in this manner.

Unfortunately the inherent limitations of a large number of quality assessment measures renders them unsuitable for many of the above purposes. The main
characteristics that should be considered when determining an appropriate quality metric for compressed natural pictures are listed below.

- **Accuracy.** This is probably the most important characteristic of a quality measure. It must closely and monotonically reflect the picture’s quality as it would be judged by the average human viewer.

- **Speed.** A measure which produces its results quickly can aid the design process, since the results can be rapidly assessed to improve the coder. Extensive subjective testing may take weeks or months to perform, whereas some objective measures may be performed in real time, thus enabling their use in an adaptive coding algorithm.

- **Various (multi-dimensional) output formats.** Many metrics rate a picture’s quality in terms of a single number (e.g. mean opinion score (MOS), MSE). However this is not very instructive, in that it does not give the designer of the coding algorithm any idea where the major perceptible errors are occurring, thus it does not provide useful feedback to the design process. A more instructive measure would provide several output formats (e.g. error maps, indicators of the type of distortion), and may identify particular areas or scene combinations where the coder’s perceptual performance is lacking. This information can then be intelligently used in the coder’s design.

- **Cost.** In a subjective measure, this involves the cost of actually running the tests, while in an objective measure it refers to the cost of programming (or purchasing) the quality algorithm.

- **Repeatability.** It is important that the quality measure’s predictions are repeatable and reproducible with the same accuracy at different locations and times.

- **Robustness / consistency.** The quality metric should give an accurate measure of a picture’s quality under a variety of different viewing conditions,
scene types, coders, and compression ratios. It should not fail dramatically for particular subsets of inputs.

- *Simplicity*. This is also generally related to cost and speed. All else being equal, a quality metric with lower complexity will generally be preferred.

- *Standards conformance / acceptability*. Recognised standards exist for subjective testing of quality. Although no formal standards exist yet for objective measures, various standardisation processes are currently occurring [81].

The primary purpose of this chapter is to review the different ways in which the quality of compressed pictures can be evaluated. Section 3.1 briefly reviews the various compression techniques which have been developed, and gives a summary of the types of artifacts which they produce. Techniques for assessing picture quality subjectively are discussed in Section 3.2, focusing on ITU-R Rec. 500 and its features and problems. This is followed by a review of objective quality assessment algorithms in Section 3.3, analysing the strengths and weaknesses of different approaches.

### 3.1 Compression Techniques and their Artifacts

The ability to compress natural images stems from the inherent redundancy contained in them. In general, there are three types of redundancy that can be identified in digital pictures:

- *spatial redundancy*, which is due to the correlation (or dependence) between neighbouring pixel values,

- *temporal redundancy*, which is due to the correlation between different frames in a sequence of video images,
Figure 3.1: General block diagram of lossy compression algorithms.

- perceptual redundancy, which occurs because the HVS does not require as much information as is contained in the digital representation; thus a reduction in information may not be perceived by the human viewer.

Compression can be viewed as the removal of this superfluous information, which consequently results in a reduction in bit-rate.

Lossy techniques are used in practically all video compression algorithms, and in most image compression schemes designed for general purpose viewing. Lossy compression techniques allow degradations to appear in the reconstructed signal, although they may not be visually apparent. In exchange for this, they provide a substantial reduction in bit-rate in comparison to lossless schemes. Compression ratios for typical high quality lossy images can vary from around 4:1 to over 40:1, depending on the complexity of the image, the application, and the required quality. Compression ratios are much higher for video, due to additional temporal redundancy. The general block diagram for a lossy scheme is shown in Figure 3.1. Most schemes contain all components, although simple techniques may omit some stages. The decomposition or transformation is performed in order to reduce the dynamic range of the signal, to eliminate redundant information, or to provide a representation which can be more efficiently coded. This stage is generally lossless and reversible. The losses (and most of the compression) are introduced in the quantisation stage. Here, the number of possible output symbols is reduced. Thus it is by changing the amount of quantisation that the bit-rate and quality of the
reconstructed signal is adjusted. It is important to quantise the data in such a way that the symbol encoding may be efficiently performed. This encoding stage may employ techniques such as Huffman coding or arithmetic coding, to code the symbols compactly at bit-rates close to their entropy.

### 3.1.1 Image Compression

Although image compression algorithms have been researched for a number of years, the area continues to be one of active research. Many different algorithms have been developed, with each one introducing different types of distortion. It is beyond the scope of this thesis to give a detailed analysis of the operation of the various compression algorithms. However a brief summary of some of the most common image compression techniques is appropriate. Full details of various compression techniques can be obtained from a number of different textbooks (e.g. [105, 170, 177, 200, 209]).

- **Discrete Cosine Transform-based compression.** The DCT has been preferred to the Discrete Fourier Transform and Karhuenen-Loeve Transform due to its symmetric frequency representation (resulting in reduced error at block edges) and its image independence. The image is typically broken into $8 \times 8$ or $16 \times 16$ blocks, each of which are transformed using a forward 2-D DCT. Quantisation is then applied, and the resultant symbols can be efficiently coded as they contain a large number of small values (usually zeros) at high frequencies. The DCT forms the basis of the JPEG, MPEG, and H.261 compression standards [147, 170, 200] (see Section 3.1.3).

- **Wavelet compression.** Wavelet-based compression schemes [232] have recently become an area of active research. One of the main reasons for this interest is because wavelets provide localisation in both space and spatial
frequency, which is similar to the operation of neurons in the primary visual cortex. Because the technique is not block-based, wavelet coded images typically degrade gracefully at low bit-rates.

- **Subband coding.** Subband coding [286] is related to wavelet coding. However, rather than applying octave-width frequency bands, subband coders typically use filters which are uniformly spaced in spatial frequency. The different subbands can be quantised separately, in accordance with human perception.

- **Vector Quantisation (VQ) [105].** Rather than quantising individual pixels, VQ operates by quantising small groups of pixels (typically $2 \times 2$ or $4 \times 4$ blocks). A codebook must be generated by both the coder and decoder, and this codebook contains a list of codevectors which give a representative sample of the pixel groups typically found in a scene. The size of the codebook is much lower than the total number of pixel combinations possible, so by transmitting only the codevector index number, significant compression occurs.

- **Fractal coding [86].** Although heavily researched in the late 1980s and early 1990s, fractal coding has recently received reduced attention. The basic idea is similar to VQ: limited subsets of image blocks are used to closely estimate the blocks in a natural image. However, fractal compression also allows affine transforms to occur on the blocks. Although decompression is fast, fractal encoding is computationally very expensive.

- **Region-based compression.** Region based methods [137] segment the image into uniform regions surrounded by contours which correspond, as closely as possible, to those of the objects in the image. The contour and texture information is then coded separately. Contours may be extracted in two ways: by using edge detection techniques, or by region growing. The textures contained within each of the regions are generally quite uniform,
and can be efficiently coded using a variety of methods such as polynomial coding or shape-adaptive DCT.

- Model-based coding. In model-based methods [4] the coder and decoder agree on the basic model of the image, so that the coder then only needs to send individual image parameters to manipulate the model. This relatively new method promises extremely low bit rates, but the types of applications may be restricted to those with simple scenes (e.g., head and shoulders, as in videophone applications), where sufficient a priori information about the structure of the scene is known.

### 3.1.2 Video Compression

The compression of sequences of images (video) will generally result in lower bit-rates per picture than still images, due to the presence of temporal redundancy. Video compression algorithms employ a variety of techniques aimed at removing this redundancy. The choice of algorithm will depend heavily on the nature of the compression method being used (e.g., block-based, model-based, pixel-based or region-based), since each of these methods will require a different way of removing temporal correlation.

The simplest way to reduce temporal redundancy is to code the difference between two frames rather than the frame itself. Since there is a high correlation between frames, the difference image should be small, except in regions of movement. The motion in the image can be identified using a motion estimation algorithm. In general, the motion of objects in a scene may be very complex. However, most current estimation and compensation techniques assume that the only motion is a simple translation in a plane that is parallel to the camera plane, since this allows a computationally feasible solution. Block matching motion estimation algorithms are currently widely used. These techniques assume that the object
displacement is constant within a small block of pixels (e.g. $8 \times 8$, $16 \times 16$). Thus the aim of such algorithms is to find the motion vector of each pixel block, since the actual pixel values from one frame to the next are likely to be very similar. Only the motion vector (and perhaps the difference between the blocks) will need to be coded, rather than the entire block contents. A full review of motion estimation and compensation techniques is contained in Mitchell et al. [170].

### 3.1.3 Compression Standards

Compression standards have provided a mechanism for interoperability between different applications which utilise compression. At the same time, the standards specifications have been flexible enough to allow proprietary compression techniques to be used, since only the coded bit-stream is defined. This section briefly describes the common image and video compression standards, and outlines future standardisation processes. Detailed descriptions of the standards are contained in the specified references.

#### 3.1.3.1 JPEG

The JPEG compression standard [200, 266] was released in 1991, and was produced as the culmination of a lengthy process which combined the best features from a number of different techniques. Four different schemes are specified within the standard: sequential, progressive, lossless, and hierarchical. However, the baseline sequential algorithm is the one which is most commonly implemented. The baseline sequential algorithm involves the following steps: a decomposition of the image into $8 \times 8$ blocks, a forward-DCT on each block, quantisation using a coder-selectable quantisation table, and run-length and entropy coding of the zig-zagged frequency co-efficients. The decoder performs the reverse processes to recover the original image. Currently a new standardisation effort is occurring
which is termed JPEG-2000. Algorithms for this standard were still being assessed at the time this thesis was written, but it is likely that techniques other than DCT-based will be accepted in this standard.

3.1.3.2 MPEG

The original MPEG-1 video coding standard [139] was released in 1992 and was aimed at multimedia applications (up to 1.5 Mbit/sec). It was strongly influenced by JPEG and is consequently DCT-based. Motion compensation is performed via block estimation, and frames are transmitted as either intra-pictures (I), predicted pictures (P), or bi-directionally interpolated pictures (B). MPEG-2 [170] was released in 1994 and was aimed for use in broadcast television. It is very closely related to MPEG-1, but includes various additional features such as support for interlacing and more flexible picture profiles.

A new MPEG-4 standard [208] is currently in the final stages of standardisation. It was initially intended for very low bit-rate applications (≤ 64 kbit/sec), but its scope has since extended to also include higher bit-rates, interactivity, and communication across computer networks. The specification of the standard is very open, such that a variety of compression techniques may be used. A significant feature of MPEG-4 is that it allows for the coding of objects in a scene. Therefore, different objects in a scene can be coded at different compression levels, depending upon the significance of that object within the scene. This object-based structure, and the associated spatial and temporal scalability which is possible at the object level, presents a significant challenge for quality assessment algorithms.

Work has recently started on a new MPEG-7 standard. This technique is aimed at content based retrieval of audiovisual information. In this sense, it is independent from the previous MPEG standards. MPEG-7 is expected to be submitted as an international standard by the end of 2001.
3.1.3.3 H.26x

The H.261 video coding standard [147] was released in 1990, and was designed for use in visual telephony. Its bit-rate is therefore quite low, being \( p \times 64 \) kbit/sec \((p = 1, 2, ..., 30)\). H.261 can basically be considered as a simplified version of MPEG-1 (i.e. block-based DCT with motion compensation). In 1995, a more advanced version of H.261 called H.263 was standardised. The basic operation is generally quite similar to H.261, but various improvements such as increased bit-rate flexibility, half-pixel-accuracy motion compensation, and unrestricted motion vectors have been adopted. These standards form part of the H.320 standard for audio-visual services at low bit-rates. H.320 includes not only video coding but also audio coding, network multiplexing, and signalling standards.

3.1.4 Artifacts Introduced by Digital Compression

The lossy compression techniques discussed in this chapter introduce a large number of different artifacts into the reconstructed picture. The effects that these distortions have on human perception of picture quality vary significantly and depend on factors such as the artifact’s magnitude, spatial and temporal persistence, location, and structure. The following list (modified from [61, 292]) describes many of the most common coding artifacts, together with a summary of the compression schemes where each particular artifact is likely to occur.

- **Basis image blocks**: reconstructed blocks have resemblance to basis blocks of transformed image due to severe quantisation. Can occur in block-based transform coders.

- **Blocking / tiling**: distortion characterised by the appearance of discontinuities at the boundaries of adjacent blocks. Produced by block-based transform coding (e.g. JPEG, MPEG) and VQ.
- **Blurring**: a global distortion over the entire image, characterised by reduced sharpness of edges and spatial detail. Found in most lossy schemes.

- **Colour errors**: distortion of all or a portion of the final image characterised by the appearance of unnatural or unexpected hues or saturation levels. Occurs in most lossy schemes.

- **Edge busyness**: distortion concentrated at or near the edge of objects. May be either temporal (time varying sharpness or shimmering), or spatial (spatially varying distortion near edges). Produced by most lossy schemes.

- **Error blocks**: distortion where one or more blocks bear no resemblance to the current or previous scene and often contrast greatly with adjacent blocks. Can occur in any block-based scheme undergoing channel errors.

- **Granular noise**: random or granular noise structure appearing in flat areas of the picture. Produced by differential pulse code modulation.

- **Jerkiness**: motion that was originally smooth and continuous is perceived as a series of distinct “snapshots”. Any video coder which uses temporal sampling or interpolation can introduce jerkiness.

- **Motion-related artifacts**: noticeable distortion in motion areas, which generally increases in areas of higher motion. Can occur in any lossy video compression scheme.

- **Mosquito noise**: a form of edge busyness associated with movement, characterised by moving artifacts and/or blotchy noise patterns superimposed over objects. Occurs in transform coded images.

- **Object persistence**: distortion where objects that appeared in a previous frame (and should no longer appear) remain in current and subsequent frames as an outline or faded image. Can occur if temporal interpolation is carried out over too many frames.
- **Ringing**: additional contouring around edges, resulting in the appearance of edge-like artifacts extending out from the edge. Common in subband, wavelet, and DCT-based coding.

- **Smearing**: a localised distortion over a sub-region of the image, characterised by reduced sharpness of edges and spatial detail (e.g., a fast moving object may exhibit smearing). Most lossy video compression schemes can introduce smearing.

Although many of these artifacts would only become visible at low bit-rates, it can be appreciated that there are a large number of different artifacts that may be introduced as a result of coding. It is therefore impractical to design a quality metric which is aimed at detecting each artifact separately, since it would require prior knowledge of the type of coding method that was used. In addition, it may not be possible to separate the different artifacts, as many of them may be correlated. A more complete model of perceptual distortion is desirable: one which gives accurate results regardless of the type of coding used on the image. Such a measure would also be able to detect any new artifacts that may be introduced by future coding algorithms.

### 3.2 Subjective Quality Assessment Techniques

Subjective methods of quality assessment involve the use of human subjects in evaluating the quality of a coded picture. ITU-R Recommendation BT.500-6[166] is a commonly used standard for the subjective assessment of still images and video. Four assessment methods are recommended in the standard, with the choice of which particular method to use being left to the experimenter, depending on the purpose of the test. These methods are summarised below.
• **Double stimulus impairment scale method.** This method is intended for images or sequences which cover a wide range of impairments. Subjects are shown the original picture followed by the impaired picture, and are asked to rate the quality of the impaired picture with respect to the original. Results are indicated on a discrete five-grade impairment scale: imperceptible, perceptible but not annoying, slightly annoying, annoying, and very annoying.

• **Double stimulus continuous quality scale (DSCQS) method.** This method involves two images or sequences, one of which is a reference. However, the subject is not told which one is the reference and is asked to rate each picture independently. Results are marked on a continuous scale which is divided into five sections. The sections are marked with adjectives for guidance: excellent, good, fair, poor, and bad. The continuous quality scale allows greater precision in judgments to be made by the observers. This kind of test is best suited to situations where the impairments are relatively small.

• **Single stimulus methods.** These techniques are useful when the effect of one or more factors needs to be assessed. The factors can either be tested separately or can be combined to test interactions. Subjects are shown the test image or sequence and are asked to rate its quality. A number of different ways of recording observer results are possible. These include: a discrete five-grade impairment scale (as per the double stimulus impairment scale method); an 11-grade numerical categorical scale; continuous scaling, where the subject indicates the score on a line drawn between two semantic labels; and numerical scaling, where subjects are free to choose their own numerical scale, thus avoiding scale boundary effects.

• **Stimulus comparison methods.** When two impaired images or sequences are required to be compared directly, these techniques are most useful. Subjects
are shown the two distorted scenes in a random order and are asked to rate the quality of one scene with respect to the other. Subjects’ opinions are typically recorded using adjectival categorical judgments, where the subject compares the scenes using a discrete seven level scale (much worse, worse, slightly worse, the same, slightly better, better, and much better). Continuous scales or performance methods may also be used.

The Rec. BT.500-6 standard also specifies several other features to be considered in the testing, which are listed below.

- **Number of subjects**: at least 15, and preferably more. They should have normal or corrected-to-normal vision, and preferably should be non-expert.

- **Test scenes**: these should be critical to the impairment being tested. Although useful results can still be obtained using only two different pictures, a minimum of four is recommended for most tests.

- **Viewing conditions**: specifications have been established for the room environment, monitor, ambient lighting conditions, viewing distance, and viewing angle.

- **Stimulus presentation**: random presentation of sequences is recommended, with no sequence being displayed twice in succession. Test sessions should not last longer than half an hour.

- **Data processing**: after subjective testing has been performed, the data obtained from the experiments may be further analysed. Techniques employed in this process include averaging over several samples, removal of a subject’s mean score from their rating, and removal of outlier data. Following such processes (and if the sampling size is high enough) an accurate measure of a picture’s quality may be obtained. In a study of 114 subjects by Cermak
Subjective testing is currently the accepted method for accurately establishing the quality of a particular coding technique. However, there are several problems associated with performing subjective quality tests. The list below summarises the most important of these. Some are related to specific types of subjective tests, and their effects may be reduced or eliminated through the use of appropriate testing and post-processing procedures. However, many of these problems are inherent to subjective testing. It is these intrinsic difficulties which have driven research into objective quality assessment techniques.

- Subjective tests are extremely time consuming and costly. Many groups that research compression techniques do not possess the required laboratory viewing conditions, and must either perform their testing at other locations or conduct tests under non-standard conditions. It is also difficult to obtain a large number of subjects to perform the tests. The process of subjective testing may take in the order of weeks or months, thus becoming a major bottleneck in the development of a coder.

- In order to obtain accurate results, a large number of subjects are required. This is because there can be a large variation in individual viewing opinions, depending on the subject’s age, sex, experience, motivation, or other personal factors. Cermak and Fay [34] used 114 subjects and, under the Rec. 500 conditions and using various data analysis techniques, were able to reduce the standard error for the average coder–scene combination to 0.11 scale points.

- Subjective test data is only valid for the particular set of viewing conditions that the test was performed under. New tests need to be done to extend the results to other conditions.
- Subject’s quality assessments may vary considerably depending on the length of the sequence and the location of the major distortions. Aldridge et al. [5] performed subjective tests using 30 second sequences, and found that subjects gave significantly more weight to impairments occurring towards the end of the sequence (up to half a scale point on a five-point scale). This is termed the *recency effect*. Hamberg and de Ridder [51, 102, 103] have developed a model which predicts subject’s overall MOS rating from the dynamic changes in quality occurring during a sequence. By measuring subjective quality instantaneously, they also found that subjects take around 1 sec to react to a particular distortion in a scene, and a further 2–3 sec to stabilise their response.

- The scale used by subjects to record their rating can also introduce problems. Discrete scales with only a few levels automatically introduce significant approximation, thus requiring a large number of subjects to reduce variance. If a fixed scale is used, subjects are typically reluctant to assign very high or very low scores, as this would then leave no room to rate subsequent scenes with very high or very poor quality. Several techniques such as anchored scales and unrestricted numerical scaling may alleviate this problem.

- The results of subjective tests generally give little or no indication as to *where* (either spatially or temporally) the errors occurred in the pictures. Therefore they provide limited feedback to the designer of the coding algorithm as to the *location* of major problems. Instantaneous subjective quality ratings like those suggested by Hamberg and de Ridder [102] address the temporal aspect by recording subjective quality throughout the duration of the sequence.

- Van Dijk et al. [257] report that when two different coders are being tested, subjects have difficulty rating quality *across* the different coders and instead
use a separate scale for the separate coders. A similar effect can also occur across scenes. Recently a double anchored numerical category scale technique has been proposed by Avadhanam and Algazi [12] which addresses this problem. Using this technique, subjective tests are first performed within particular coders and scenes, and then compared across coders and scenes.

A final point to consider is that, although subjective tests are currently the most accurate method of assessing picture quality, their results are not perfect. Even the best subjective test data will still have some degree of error or variance. Objective techniques which use subjective data as ground truths to assess the accuracy of their technique should take this into account; perfect correlation with subjective quality data may not be realistically attainable or desirable.

3.3 Objective Quality Assessment Techniques

Objective methods of picture quality evaluation (i.e. mathematical algorithms which are designed to correlate well with human viewer response) have been developed as a result of the various problems associated with subjective testing. This chapter reviews the different types of objective metrics which have been developed. Although many approaches have been proposed, the metrics which are still most widely used are the simple MSE and PSNR, even though the inaccuracies of these techniques have been well documented [44, 142, 257, 289]. This unfortunate situation has occurred due to a lack of accurate objective quality metrics, and a lack of standardisation. However, in recent years these two issues have started to be addressed. Active research into objective picture quality evaluation techniques has seen a significant improvement in the accuracy and robustness of proposed algorithms. In addition, standardisation attempts by ANSI [60, 61] and
more recently by the ITU Video Quality Experts Group (VQEG) [81] give promise that effective alternatives to PSNR and MSE may soon be widely adopted.

### 3.3.1 PSNR, MSE, and Variants

The simplest and still the most widely used objective quality assessment metrics are MSE and PSNR. These are defined as

\[
MSE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (x_{ij} - \hat{x}_{ij})^2; \tag{3.1}
\]

and

\[
PSNR = 10 \log \left( \frac{(2^n - 1)^2}{MSE} \right) dB; \tag{3.2}
\]

where \( M, N \) are the number of horizontal and vertical pixels, \( x_{ij} \) is the value of the original pixel at position \((i, j)\), \( \hat{x}_{ij} \) is the value of the distorted pixel at position \((i, j)\), and \( n \) is the number of bits per pixel. The use of these measures in picture quality assessment probably stems back to the evaluation of error in other areas of communication. In many other areas, these simple measures are indeed appropriate and give an accurate measure of a picture’s quality. They continue to be commonly used in picture quality evaluation primarily due to their speed and simplicity, and because of a lack of any other standard or globally accepted objective measures. However the shortcomings of these measures are also widely recognised [44, 142, 257, 289]. Although PSNR and MSE give a monotonic indication of quality for a particular scene and coder, they perform poorly when tested across different scenes and coders. This is because they treat all numerical impairments as being of equal importance, regardless of their location in the scene. These types of metric fail to take into account any of the properties of the HVS discussed in Chapter 2, and therefore cannot perform
robustly across scenes and coders. Tests on typical scenes show that the PSNR and MSE show a correlation with subjective MOS data of only $r = 0.4\pm 0.7$.

Several simple improvements on these techniques have been proposed. These include weighted signal-to-noise ratio (WSNR) [106] and normalised mean square error (NMSE) [167, 162]. These techniques crudely try to model some HVS properties such as the CSF or spatial masking. Although they offer improved correlation with subjective opinion, these techniques are still too simple to provide sufficient accuracy and robustness over a wide range of scene types and coders.

### 3.3.2 Correlation-based Techniques

A group of quality models have been designed which generate a measure of quality by first extracting several parameters from the original and encoded images (or the error image formed by the difference between the original and encoded images). These parameters are chosen since they are believed to be related to the viewer’s perception of a coded picture’s quality. Typical parameters which are chosen include blurriness, blockiness, noisiness, and temporal features such as jerkiness in video metrics. Subjective quality assessment data is obtained, preferably from a wide range of scenes, bit-rates, and coders. The correlation between the extracted parameters and the subjective test data is determined via a regression analysis or a neural network. A predictor of image quality is then obtained based on a weighting of the parameters.

A substantial effort has been made by ANSI to establish a standard for objective video quality using the correlation-based approach. The project began in 1989 and recently resulted in the release of a draft standard [60]. This standard proposes various different parameters that may be extracted from the original and coded sequences, but does not elaborate on how the parameters should be best used to predict the quality of a sequence. However, earlier papers published by
the group (e.g. [273]) proposed ways to implement the quality measure. Three parameters were extracted from the original and coded sequences: one relating to the amount of spatial blurring, and two temporal features relating to the amount of missing motion and added motion respectively (thus reflecting the temporal blurring and jerkiness in the sequence). These parameters were chosen firstly since they were highly correlated with subjective quality, and secondly because they only displayed limited correlation with each other. A linear predictor of the picture’s quality was formed based on a regression with subjective tests using a least square criterion. The results reported by the group are quite good, and a correlation with subjective MOS data of up to $r = 0.94$ has been reported. However, other researchers (e.g. [146, 256]) have found that this metric correlates quite poorly with subjective opinion, particularly when tested over a wide range of bit-rates. A number of metrics which use an approach similar to the ANSI metric have recently been developed (e.g. [43, 91]). Although reasonable correlation with subjective opinion is reported over a limited range of scenes, coders, and bit-rates, these techniques are all reliant on the choice of subjective test data set and therefore are not robust over a wide range of input data.

### 3.3.2.1 Hybrid HVS/Correlation-based Techniques

Several quality prediction models have been developed which firstly perform some HVS-based processing, then extract various parameters from the processed image, and finally produce a predictor of quality based on the correlation between these parameters and subjective quality rating data. The first such technique was Miyahara’s *picture quality scale* (PQS) for the assessment of image quality [171]. This algorithm performs brightness non-linearity and CSF processes on the error image, and then conducts a principal component analysis to extract parameters relating to random errors, blocking artifacts, and errors along contours. A linear predictor based on these parameters is then obtained via a multivariate analysis
with subjective MOS data. Results showed quite good correlation with MOS but only a very limited data set was used. Since the original PQS was proposed, a number of papers have added improvements, including extensions to video [101, 116] and inclusion of masking and structural error parameters [11].

Xu and Hauske [289] include a relatively complex visual model in their algorithm. This includes a brightness non-linearity, CSF, and masking processes. They then extract parameters relating to errors on ‘own edges’, ‘exotic edges’, and ‘error in other areas’. A simple quality predictor is then formed, based on these parameters. The results of their model show an average correlation with subjective tests of $r = 0.875$, a substantial improvement over PSNR ($r = 0.653$ correlation). However, the most significant improvements over PSNR occurred in unnatural test scenes; in natural scenes the difference between the two measures was substantially smaller. Several other techniques have also adopted a similar hybrid HVS/correlation-based approach [35, 47, 184].

### 3.3.2.2 Problems of Correlation-based Techniques

Most of the correlation-based techniques report quite good correlation (around $r = 0.90$) with subjective MOS data, over a limited range of scenes, bit-rates, coders, and viewing conditions. Therefore, for applications where these factors remain restricted, quite good correlation can be achieved. However, the inherent problem with these techniques is that they are constrained to operate accurately only for a range of viewing material similar to that used in the subjective quality rating test from which they were calibrated. Therefore, if different scenes, bit-rates, coders, or viewing conditions are used, the usefulness of these techniques for quality prediction can no longer be guaranteed. A more comprehensive and wide-ranging subjective test database could be used, but this would reduce the accuracy achieved by the technique within the range of test conditions used. In addition, correlation-based techniques rely heavily on the accuracy of the subjective test
data. As discussed in Section 3.2, the accuracy of subjective data cannot be taken for granted, and there will always be some inherent variability. The reliance on the subjective test data is therefore a severe restriction on the robustness and general applicability of correlation-based techniques.

Another problem of correlation-based techniques is that they typically only produce a single number representing overall picture quality. As discussed earlier in this chapter, this is not very helpful for the designer of a coding algorithm, since no indication of the location of coding errors is given.

3.3.3 HVS-based Techniques

The inherent problems associated with other objective techniques, coupled with the improved understanding of the operation of the HVS, has resulted in the development of quality assessment techniques based on models of the HVS. The logic behind such an approach is obvious: the quality of a picture is ultimately judged by the HVS, so if it is possible to accurately model its operation, then the resulting quality algorithm should give an accurate prediction of the picture’s quality. As can be seen in Chapter 2, the task of modeling the operation of the HVS is a difficult one. Early attempts at incorporating simple visual properties into a quality assessment algorithm [28, 100, 160, 220] generally achieved only moderate improvements over MSE in terms of correlation with subjective opinion. However, more recent models which are based on the operation of the HVS (e.g. [44, 152, 256]) have been shown to accurately measure image fidelity, which indicates that HVS-based techniques may be suitable for use as robust, flexible quality metrics.

HVS-based quality metrics offer several other advantages. In addition to producing a single number for correlation with subjective MOS, HVS-models can produce a visible error map, which indicates exactly where the perceptible errors
are located in the scene. This map can be further analysed to determine the exact nature of the distortion. Perhaps the most appealing feature of HVS-based models is their robustness. They are not intrinsically restricted to any particular type of scene, distortion, or bit-rate, which makes them generally applicable. Viewing parameters are typically used as input parameters to the model, so different viewing distances, monitors and ambient lighting conditions can be tested simply by adjusting these input parameters. As with all objective techniques, the results are readily repeatable. Although quite computationally complex, modern processors enable results to be generated in real-time or close to real-time.

Despite the many attractive features that HVS-based quality metrics possess, there are still many areas in which they can be improved. A full review of contemporary work in HVS-based models is given in Chapter 4, along with a detailed discussion of current problems. The techniques demonstrated in Chapters 5 and 6 of this thesis directly address these deficiencies, and in Chapter 7 a new HVS-based metric for assessing the quality of compressed natural images is proposed.

### 3.3.4 Other Techniques

A number of techniques have been developed which cannot be easily categorised, and so are described separately in this miscellaneous collection. Many of these techniques are aimed at specific applications. Nill and Bouzas [181] developed a technique which could be useful in situations where no reference scene is available. Their technique is based on the finding that natural scenes contain a power spectrum which drops off as an inverse square with spatial frequency [30, 62, 84, 245]. Compression algorithms distort this spectrum (typically by reducing high frequency components), so Nill’s technique examines the power spectrum of the coded scene to estimate the level of compression which has occurred. Although the technique may not be applicable for all types of scenes, it has obtained good
results for aerial images (which generally follow the inverse-square power spectrum rule closely and reliably).

Van den Branden et al. [255] used synthetic test patterns to test the quality of a digital coder. In this technique, both the transmitter and receiver contain a test pattern generator. The transmitter indicates which patterns it is going to transmit by first sending a synchronisation frame in which the information is coded. It then sends the test patterns. The receiver determines the test patterns which have been sent from the synchronisation frame. The original patterns are then generated using the receiver’s pattern generator. Any of a number of techniques can then be used to assess the quality of the coded test patterns. This approach is also useful in situations where the original scene is not available to the receiver (e.g. broadcast television). However, it is doubtful whether test patterns can representatively sample the range of viewing material presented to the coder [300], so such approaches may not fully test the capabilities of the coder.

Other application-specific quality metrics which have been proposed include Baddeley’s metric for binary images [13], a magnetic resonance imaging (MRI) quality metric by Anderson et al. [9], and a model for aerial images by Boberg [20]. However, these techniques are designed for a particular purpose, and are unlikely to perform well for typical natural images. Rather than adding to the myriad of techniques which have been proposed, it would be desirable to have a robust, general framework on which a variety of (perhaps application specific) quality metrics could be developed. Early vision models are strong candidates for providing such a framework, and are reviewed in Chapter 4.
Chapter 4

Objective Quality Assessment Based on HVS Models

In Chapter 3 HVS-based models were identified as being a promising technique for assessing the quality of compressed pictures. Such models are not restricted in terms of the types of image, artifact, bit-rate, or viewing conditions on which they can operate. In addition, HVS-based techniques indicate exactly where the visible distortions are occurring. This information is useful feedback for the designer of the compression algorithm. Other objective techniques have been shown to work well for a particular range of inputs, but they invariably suffer from limitations which inhibit their general applicability.

Current HVS models aim to mimic the operation of the visual system up to and including the primary visual cortex. They typically consist of five main components:

- luminance to contrast conversion,
- channel decomposition,
- CSF,
• masking,

• summation.

Most models sequence the components in the same order as above, since this approximates the way that the processing occurs in the HVS. However, there are exceptions. Some models leave out either the contrast conversion, channel decomposition, CSF, or masking components altogether (e.g. [100, 160]). Other models combine components together. For example, Teo and Heeger [237, 238] combine the CSF and spatial masking processes into a single non-linearity.

A large number of different HVS models have been proposed for a variety of different vision-based tasks, such as image quality evaluation, object detection and object discrimination. These models have generally obtained their parameters from psychophysical tests of detection thresholds using simple visual stimuli. They therefore are good at predicting detection thresholds for similar simple stimuli. However, the non-linear and adaptive processing of the HVS means that such models do not necessarily hold when suprathreshold or complex stimuli are used. Recent models show promising results for object detection tasks in natural scenes [3, 74, 214]. However, considerable improvements must still be made before HVS models can accurately predict human response to a wide range of naturalistic and suprathreshold stimuli.

The focus of this chapter is to perform a critical analysis of current state-of-the-art human vision models for image quality assessment, and to determine areas where these techniques can be improved. A review of previous HVS-based models is provided in Section 4.1. This is followed in Section 4.2 by an analysis of the deficiencies which these models have, particularly with respect to the quality assessment of natural images with suprathreshold errors. In Section 4.3, some specific features of HVS models are examined and appropriate ways of implementing these features for natural image quality assessment are proposed.
4.1 Previous HVS-based Models

Numerous techniques for modeling the operation of the early stages of the HVS have been proposed. Some of the models aim to be neurophysiologically plausible: for instance, by mimicking the response of neurons in the primary visual cortex. Others perform modeling in a less direct manner, and consider the HVS as a “black box”. These techniques model overall operations or features of the HVS in a way that may differ significantly from the internal operation of the HVS, but the model’s outputs closely resembles those of visual processes.

Models of the HVS can be classified into two broad groups: *single channel* and *multiple channel*. Multiple channel models decompose the image into multiple spatial frequency and/or orientation bands, and determine thresholds separately for each channel before summation. They are therefore inspired by the operation of cortical neurons, which are known to be spatial frequency and orientation selective (see Section 2.1). Single channel models do not perform such a multi-resolution decomposition, and perform their operations globally on a single-scale representation of the image.

4.1.1 Single Channel Models

One of the earliest spatial domain models was developed by Mannos and Sakrisson [160]. This simple model consists merely of a luminance non-linearity (implemented as a cube-root function) followed by a CSF whose parameters were chosen by subjective tests. No actual analysis of the accuracy of the measure was given, but subsequent tests show that its correlation with subjective test data is quite low for JPEG encoded images [93]. This could be expected not only because of the simplicity of the model, but also because of the limited stimuli used to determine model parameters. Another model which contained only the luminance non-linearity and CSF stages was that of Hall and Hall [100]. In their model,
the CSF is implemented as separate low-pass and high-pass filters. The luminance non-linearity is applied after the low-pass filter but before the high-pass filter. This attempts to model physiological processes: the low-pass nature of the HVS is predominantly due to the optics, the brightness non-linearity occurs in the receptors, and the high-pass characteristics are caused by the centre-surround response of ganglion neurons. However the results of Hall and Hall’s model are inconclusive and this type of two-stage CSF model has not been adopted by any recent HVS models.

Limb’s image distortion criterion [142] consists of luminance non-linearity, CSF, spatial masking, and summation stages. Various ways of implementing these components were investigated, such as global and local masking, different shaped CSFs, and various summation methods. The best implementation was determined by comparing the model’s prediction with subjective MOS data. Highest correlation was found using the following configuration: a CSF which was low-pass in shape, local activity masking, and a summation model which focuses on the two or three areas of worst distortion, rather than averaging across the scene. The results show a significant improvement in correlation with MOS in comparison to the MSE metric. However, the calculation of model parameters is dependent upon the subjective test data used, so it is only optimal for similar types of scenes and viewing conditions. The model is therefore restricted in a similar way to the correlation-based techniques discussed in Section 3.3.2.

Lukas and Budrikis used a similar methodology to Limb, but their model was extended to video [154]. The spatio-temporal CSF is implemented as the division of separate excitatory and inhibitory paths, which are linear but combine in a non-linear way. Spatial masking is implemented using a simple local activity measure, and summation is performed using a Minkowski metric with an exponent of 2.0. Results show a correlation of $r = 0.88$ with subjective MOS, compared to only $r = 0.69$ for MSE. This shows that significant improvements in correlation can
be achieved, even if some visual processes are modeled in a crude or empirical fashion.

More recently, Karunasekera and Kingsbury [131] presented an image distortion model aimed at detecting blocking artifacts. The human visual model used was interesting in that the parameters were chosen from suprathreshold experiments, which measured subject’s reaction time to various suprathreshold stimuli. This is in contrast to most other models, which derive their parameters from threshold experiments, and assume that the data scales linearly into suprathreshold regions. The model includes luminance non-linearity and spatial activity masking stages, followed by a simple error averaging around block edges to determine the magnitude of blocking artifacts. The algorithm has also been extended to include ringing and blurring artifacts [130]. The results in both of these cases show a good correlation with subjective rankings, although the test set used was very limited (only one coder and one image tested). A problem again inherent with such techniques is that they are dependent upon the type of artifact present in the coded image, and therefore cannot be generally applied.

Numerous other single channel quality models have been proposed (e.g. [97, 148, 151, 219]), which follow similar methodologies to the techniques discussed earlier in this section. Although the correlation with subjective MOS reported by these models is improved in comparison to MSE, none of these single channel models have been shown to work accurately over a wide range of different scene types, coders, and bit-rates.

### 4.1.2 Multiple Channel Models

Multiple channel early vision models have recently gained popularity. These models are based on the response of neurons in the primary visual cortex, which are known to be selective in spatial frequency (average bandwidth 1.0–1.4 octaves)
and orientation (average bandwidth 35–40 deg). Multiple channel models decompose the image into channels which are sensitive to both spatial frequency and orientation, and then determine detection thresholds individually for each channel, before summing across channels to produce a map showing the location of visible errors. Various studies have shown that multi-channel models are capable of predicting the visibility of patterns which cannot be predicted by equivalent single channel models [63, 82].

The first multi-channel image quality metric was proposed by Sakrison [220]. This technique first performs a logarithmic luminance non-linearity on the error signal, and then decomposes the resultant signal into a number of spatial frequency and orientation selective channels. CSF and spatial masking data, taken from psychophysical tests using simple stimuli, are used to determine the threshold of detection for each channel. Following thresholding, the errors are summed across channels to produce the final visible distortion map, and may be summed spatially to produce a single number representing error. When this model was proposed in 1977, the limited processing power available meant that the model was unfortunately not computationally tractable. Consequently, no testing with subjective data was reported, so it is difficult to assess the accuracy of this model.

Over a decade passed before multi-channel vision models were again considered in the image processing community for image quality evaluation. However, vision researchers continued to actively pursue the multi-channel approach. Useful vision modeling tools such as the cortex transform of Watson [269] were used as the basis of the many multi-channel image quality evaluation models proposed since the late 1980s. One of the first of these was proposed by Zetzche and Hauske [297]. They used a ratio of Gaussian (ROG) pyramid followed by orientation-selective Gabor filters to decompose both the original and coded scenes into a set of spatial frequency and orientation selective channels. Five resolution levels and six orientations were used, giving a total of 30 channels. Threshold elevation (see
Section 2.2.3) was determined from a set of masking experiments which were performed on a range of edges and lines. This threshold value was compared to the difference between the original and coded signals for each channel, to determine whether the error is visible in that channel. Summation was performed across channels and spatially to produce a single number intended to correlate well with picture quality. Correlation with subjective MOS for a large stimulus set (8 images, 13 distortions, 5 subjects) showed quite poor correlation ($r = 0.61$), although this was an improvement over PSNR which only had a correlation of $r = 0.46$. The disappointing correlation achieved by this technique could have been caused by the heuristic way in which model parameters were chosen.

The Visible Differences Predictor (VDP) algorithm of Scott Daly [44, 46] is one of the best known models for image fidelity assessment. An important feature of this model is that it firstly calibrates for specific viewing parameters (e.g., pixel spacing, viewing distance, monitor characteristics). The original and distorted image luminance values are passed through a luminance non-linearity (to account for the Weber effect) and a robust CSF. A modified version of Watson’s cortex transform [269] is then used to decompose the images into multiple frequency and orientation sensitive channels. The masking process is implemented as a threshold elevation, which adjusts each channel’s visibility threshold (from the CSF) depending on the contrast within the channel. Both the original and distorted images are used as maskers in a process which is termed mutual masking. This is performed because the error image can be masked by either the original or distorted image, depending on the type of distortion that occurs. For instance, a blurring error should be masked using the distorted image, since it is the distorted edge that the viewer is being masked by and not the original. However, in the case of blocking errors, the original image should be used as the masker, since if the distorted image was used as the masker, then the error would effectively be masking itself. Once the difference between the original and coded pictures has been compared to the threshold, the errors are combined via probability summa-
tion to produce a JND map. This indicates the probability of detecting the error at each location in the image. Such a map is useful since it shows the shape and location of errors to the designer of the coding algorithm, and is amenable to usage in an adaptive coder. However, the problem with such an output format is that it is difficult to verify the accuracy of the JND map for natural images. To test its ability to detect thresholds, Daly [45] suggests using the peak value in the VDP map as an indicator of threshold visibility. Using this approach, Daly showed that his model could predict data from a range of psychophysical experiments quite well. However, Daly did not verify the ability of the model to predict natural image quality.

Another well known early vision model is that of Lubin [152, 153]. This model once again calibrates for viewing parameters. Following optics and sampling processes, the raw luminance signal is converted to local contrast using a technique similar to Peli’s LBC algorithm [195] (see Section 2.2.1 for a description of LBC). This effectively performs both the luminance-to-contrast conversion and the spatial frequency decomposition in one step. Orientation selectivity is implemented using Hilbert pairs at four orientations. Contrast sensitivity and spatial masking data is taken from psychophysical data in the literature, which tested detection thresholds for simple stimuli. A sigmoid non-linearity is used to model the masking effect (intra-channel), as this is capable of simulating the "dipper" effect found when mask and stimulus are similar. Following thresholding, summation across channels (using a Minkowski metric with exponent 2.4) produces a JND map as output. By taking the peak output of the JND map, Lubin showed that his model could predict the detectability of a range of simple stimuli. To assess its ability to predict compressed picture quality, Lubin summed across the JND map to produce a single quality figure which could be compared to subjective MOS data. Results showed a good correlation between MOS and the quality predicted by this technique ($r = 0.94$ compared to MSE’s $r = 0.81$), although the subjective set used was quite limited (4 aerial images, 1 coder, 8 subjects).
Another interesting point is that at the point where the subjective quality of the picture departs from perfect, Lubin’s model already predicts a visible error of 3 JNDs. This suggests that the calibration of the model may be too sensitive for compressed natural images, and that the complex nature of such images provides a higher tolerance of distortions. Lubin has more recently used his model for the evaluation of liquid crystal displays (LCDs) [123] and X-ray images [164].

Teo and Heeger [237, 238] proposed a model based on the response of a cat’s cortical neurons and the contrast masking data of Foley and Boynton [88]. It consists of three stages which are listed below.

- A linear transform, to decompose both the original and coded images into spatial frequency and orientation sensitive channels (implemented as a steerable pyramid [226]).

- Squaring and normalisation, which is performed to model the contrast masking. Parameters are chosen to fit the model to the data of Foley and Boynton [90].

- Detection, which involves taking the difference between the responses of original and coded channels and summation of the response over small patches of the image.

Although the model is physiologically sound, no comparison of the model’s predictions with subjective opinion was given. It is therefore difficult to assess the performance of this model for image quality evaluation purposes.

Westen et al. [275] have proposed a model which is similar in many respects to that of other multi-channel models [275]. Decomposition into channels is performed using Peli’s LBC algorithm [195], followed by fan filtering at six orientations. CSF data is taken from a model by Barten [16] and contrast masking is modeled as
a threshold elevation with a slope of 0.65 on a log-log scale. Minkowski summation is performed across channels and then spatially to produce a single number representing image quality. Comparison with subjective MOS data showed only a slight improvement in correlation over PSNR \( r = 0.84 \) compared to \( r = 0.78 \). This subjective data was taken using 6 images, 4 different coders, and 7 subjects. More recently, Westen has extended this model to video \cite{276, 277}. The interesting feature of the video model is that it attempts to take into account viewer’s SPEMs by compensating for all regions in the scene which undergo steady motion. This assumes that all steadily moving areas in a scene can be tracked by the eye, and therefore do not undergo a reduction in acuity as would be expected if only image velocity was considered. No indication of the correlation of this video quality metric with subjective MOS data has been reported.

Watson and his colleagues have produced a number of different vision tools and models for target detection, image compression, and quality evaluation. The structure of their vision models is quite similar to others which have been described in this section. Watson et al. have recently tested the ability of their models to predict object detectability in natural backgrounds \cite{3, 74, 214}. A number of different vision models were tested: both single and multiple channel, and various masking models (no masking, in-band masking, and wide-band masking). A general finding was that the masking model used was a significant factor, and that wide-band masking typically gave best results for natural images. Somewhat surprisingly, the single channel models performed just as well as the multi-channel models in many tasks.

The first multi-channel implementation of a video quality metric was reported by van den Branden \cite{253, 254, 256}. This technique involved the decomposition of the video sequence into spatially (four frequencies, four orientations) and temporally (sustained and transient) sensitive channels. The non-separable spatio-temporal CSF was modeled as the difference between separable excitation and inhibition
mechanisms, as proposed by Burbeck and Kelly [29]. CSF data was obtained from psychophysical experiments, and spatial masking was modeled as a threshold elevation with a masking gradient of 0.7 on a log-log scale. Minkowski summation was performed across all channels and spatially to produce a final estimate of distortion. Van den Branden also proposed another video quality metric [146], which is basically an extension of Teo and Heeger’s model to video. However, no indication of the correlation of either of these techniques with subjective quality rating data has been reported.

A number of other multi-channel image quality metrics have been proposed (e.g. [41, 92, 138, 236]). The underlying structure of these models is generally quite similar to that of those already discussed in this section, and so they will not be discussed in detail.

### 4.1.3 Comparison of HVS Models

Although a large number of single and multiple channel vision models have been proposed, there have been only a few attempts to compare the operation of the models to determine their suitability for different tasks. This may be the result of the high complexity and considerable number of parameters required by many of these models, which makes the task of implementing and testing several of them very cumbersome. Some models are also the subject of patents, and the full implementational details may not be publicly available. The methodology used to compare the operation of different vision models can also be quite complex. Models which predict image fidelity or image discrimination (i.e. at the threshold of visibility) can generally be tested by comparing the model’s predictions with subjective opinion of stimulus visibility, using a two-choice (visible / not visible) procedure. This process is quite straightforward for simple, artificial stimuli, but it can become difficult designing appropriate stimuli for a large range of complex natural vision tasks. On the other hand, the testing of a model’s ability to
predict image quality generally involves obtaining a database of subjective MOS data (preferably using a wide range of scenes, coders, bit-rates, and viewing conditions), and then determining the correlation between the vision model’s prediction and the subjective data. This process is therefore dependent on the accuracy of the subjective tests, and the representativeness of the test images used.

It is well known that single channel models cannot explain psychophysical test data from a number of experiments which multi-channel models can handle [63]. These include spatial frequency selective adaptation and masking effects (see [53, ch. 6] for a review). However, in a number of tests using natural stimuli, the predictions reported using single and multiple channel models are quite similar. Rohaly et al. [213, 214] compared the ability of single and multi-channel models to predict object detectability on natural backgrounds. When no contrast masking was used, the multi-channel model performed better than the single channel model. However, when a masking component was introduced, the performance achieved by both the single and multi-channel models was very similar. The multi-channel model tested here only used intra-channel masking. Surprisingly, a simple digital difference metric also performed quite well when a masking component was included. This suggests that masking may be one of the most influential components of a vision model for complex natural images.

Similar results were found by Ahumada et al. [3] for object detection in a noisy scene. In their test, a single channel model with contrast masking performed as well as a multi-channel model with masking (masking across orientation but not spatial frequency), and better than a multi-channel model without masking. The authors suggested that the choice of an appropriate masking model, which is spatially localised, is more important for target detection in natural images than the choice of single or multiple channels.

Eckstein et al. [74] tested the ability of various vision models to predict the pres-
ence of lesions in x-ray coronary angiograms. A variety of models were tested including a simple local contrast energy metric, single channel with masking, and multi-channel with a variety of masking algorithms (none, within channel, across orientation, and across both orientation and spatial frequency). Results showed that the best prediction was given by the multi-channel model with wide-band (i.e. across both orientation and spatial frequency) masking. The single channel model gave the poorest performance and multi-channel models with narrow band masking performed moderately well. The simple contrast energy and multi-channel without masking metrics gave surprisingly good results; however the authors suggested that according to more recent tests which they had performed, these metrics would not be able to perform so well.

Eckert [71] compared the ability of both wavelet- and DCT-based early vision models to predict the fidelity of compressed X-ray radiographs. Both methods provided much better prediction of fidelity than a simple signal-to-noise ratio metric. Overall, the wavelet-based model provided more consistent prediction of fidelity across different scenes than the DCT-based model. Eckert suggested that this was primarily due to the narrower masking band that was used in the DCT-based model. The DCT basis functions have a relatively narrow bandwidth, in comparison to the 1-octave bandwidth wavelet channels. Since only within-channel masking was implemented, the DCT-based model only allowed masking over a very narrow spatial frequency range; however the wavelet-based model including masking effects over a wider (1-octave) range. This paper also demonstrated the importance of using appropriate CSF data, since the models were oversensitive (i.e. predicted JND levels greater than 1.0 at the threshold of detection) due to the choice of CSF calibration data.

Some tests have also been performed which assess the abilities of different vision models to detect compression errors. Eriksson et al. [80] tested a number of different arrangements of visual models and looked at their ability to detect
the point at which distortions are just visible in compressed images. A single channel model was tested, along with a variety of multi-channel models based on the cortex transform [44, 269]. These multi-channel models used both within-band and across-orientation-channel masking approaches. Both local and global contrast measures were also tested. Subjective tests were carried out to see when observers could detect the presence of the distortion. The tests involved 6 natural images, 12 subjects, and 3 distortions (ringing, blurring, and blocking). The results showed that the best predictions were achieved by the multi-channel metrics which used across-orientation-channel masking. The single channel metrics performed well for the blurred images, but poorly for the blocking and ringing images. Global contrast metrics gave slightly better results than local contrast methods, but the area in which contrast was measured was not varied, so this result could be misleading.

Li et al. [141] compared the operation of the well known multi-channel image quality metrics of Daly and Lubin. The authors concluded that both metrics performed accurately in a qualitative sense, and that the outputs were similar in many respects. Some problems of the models were identified, such as the absence of inter-channel masking effects, and problems choosing an appropriate CSF.

4.2 Deficiencies of Previous HVS-based Models

The significant number of techniques discussed in Section 4.1 is indicative of the strong interest shown in HVS-based techniques for modeling a variety of vision tasks. Most of the models discussed were designed specifically for image quality evaluation. However, no technique has yet been shown to operate accurately over a wide range of images, coders, bit-rates, and viewing conditions. Many of these models were able to predict detection thresholds accurately for artificial stimuli. This indicates that there are two main areas which need to be addressed if these
models are to provide accurate quality predictions for compressed natural images:

- model parameters need to be chosen which accurately reflect viewer response to complex, natural scenes, rather than simple, artificial stimuli,

- higher level and cognitive factors need to be employed when converting a visible error map to a single number which represents image quality.

These two points are now examined in more detail.

### 4.2.1 Choice of Appropriate Model Parameters

Basic visual processes such as luminance to contrast conversion, frequency sensitivity, and masking have been identified as the building blocks of early vision models. A problem however is that the HVS response can change significantly depending on the type of visual stimulus to which it is subjected. Early vision models were originally developed by vision researchers for predicting visual response to simple, artificial, controllable stimuli such as sine-wave gratings and Gabor patches. These models can accurately predict human threshold response to such stimuli. However, care must be taken when applying these models to complex natural images. Some specific areas which need to be considered are outlined below.

- For simple predictable stimuli such as sine-wave gratings and Gabor patches, a global measure of contrast (i.e. contrast measured with respect to the global properties of the image) is suitable. However, for natural images such global measures are inappropriate due to the localised processing of the HVS. A localised measure of contrast is therefore important for vision models designed for natural images (see Section 2.2.1 for a discussion).
4.2 Deficiencies of Previous HVS-based Models

- As was discussed in Section 2.2.2, threshold spatial frequency sensitivity can vary significantly depending upon the nature of the stimuli shown to the subjects. It is therefore necessary to choose CSF data which has been obtained from experiments which use appropriate visual stimuli. In the case of compressed image quality evaluation, the target stimuli should resemble compression errors, and the backgrounds may vary in luminance. The shape of the CSF may also flatten as stimuli become suprathreshold (see Section 2.2.2.2).

- The large variation in performance which occurs when different masking models are used with a vision model indicates the importance of a good masking model. The spatial, spatial frequency, and orientation extents of spatial masking can be strongly influenced by the stimulus (see Section 2.2.3 for a review). Stimulus uncertainty also has a strong influence on the strength of the masking effect (see Sections 2.2.3 and 4.3).

A wide array of different filterbanks have been used in vision models to perform the multi-channel decomposition. However, it is not clear that any particular type of filterbank is more suitable for natural image quality evaluation than another. Some filterbanks have been shown to more closely resemble the response of cortical neurons than others. However, it appears that the choice of filterbank has a smaller influence on model performance than other factors discussed in this section.

4.2.2 Determining Picture Quality from Visible Distortion Maps

The area to which current HVS-based quality metrics have paid least attention is the conversion of the visible distortion map into a single number representing overall picture quality. All techniques currently use some sort of summation
across the distortion image, generally a Minkowski summation with an exponent of between 2.0 and 4.0. Although this is a reasonable first approximation [48], many important higher level visual processes are neglected by this approach. This may not be so significant in very high quality applications where the distortions are just below or at the threshold of detection. However, once the distortions go beyond the level of just being visible, higher level factors need to be considered.

The review of eye movements and attention in Section 2.3 revealed some interesting viewing patterns for natural scenes. Rather than look at all areas of a scene equally, viewers tend to return repeatedly to a small number of ROIs in a scene, even if given unlimited viewing time. When subjects view a scene in the same context (i.e. with the same instructions and motivations), the ROIs in a scene tend to be highly correlated. As we possess high acuity in only the small foveal area of viewing, our judgment of overall picture quality is strongly influenced by the quality of these few ROIs in the scene. A number of studies which varied quality across a scene depending on its content have shown this to be true [65, 96, 136, 252]. Recent coding standards such as MPEG-4 take explicit advantage of this fact by allowing objects to be defined in the scene. Different objects can then be coded with different qualities. With increasing usage of such variable-resolution coding techniques, it is important for picture quality assessment techniques not to treat all visible errors in a scene equally, but to also consider the location of the error when determining picture quality.

As well as the location of the error, its structure is also important. Errors which cause a repeatable or unnatural pattern (e.g. blockiness, ringing) are generally more objectionable than equally visible errors which are less structured, such as random noise. Distortions which add content to the scene are in general more objectionable than distortions which remove content.
4.2.3 Computational Complexity Issues

Although it is not one of the main focuses of this thesis, the model’s computational complexity must be considered for practical purposes. This is particularly true if the vision model is to form the basis of a real-time encoding algorithm, or if real-time video quality assessment is required. Some of the complete multi-channel models which have been proposed are of significant complexity. Consider a model which decomposes the image into 7 spatial frequency and 8 orientation channels. If no sub-sampling is performed, this incurs a storage and computational burden 56 times higher than if a single channel model was used. The situation becomes considerably worse once colour images or video are used. Sub-sampling can reduce this burden considerably, but aliasing issues must then be considered. Virtually all of the filterbanks used for vision modeling introduce significant redundancy following filtering, even after sub-sampling.

Techniques for simplifying vision models without seriously affecting performance are therefore important. One area which is a strong candidate for simplification is the removal of orientation selectivity in the filtering, a proposal suggested by Lubin [153] and supported by Peli [196]. Orientation selectivity is considered important for two reasons. Firstly, the CSF is known to drop slightly at oblique orientations (the oblique effect), so orientation selectivity is required for highest accuracy. Secondly, masking is generally strongest when the stimuli and masker are of the same spatial frequency and orientation. However, these factors may not be such an issue in natural images for the reasons listed below.

- The oblique effect is only present for high spatial frequencies, typically over 10 c/deg [99]. The power spectrum of natural images drops off strongly with spatial frequency, so that only a small percentage of the image’s power is contained at such high frequencies and produces such an oblique effect.

- As discussed in the next section, masking effects in natural images tend to
occur over a wide range of orientations. This is supported by the masking data of Foley and Boynton [90]. The tests of discrimination and detection models on natural images (discussed in Section 4.1.3) also give strong support for wideband masking effects, particularly across orientation channels [3, 74, 80]. Although the dipper effect cannot be modeled without orientation selectivity, its influence in natural images is likely to be minor, since most natural images consist of contrasts well beyond threshold. Therefore, orientation selectivity in masking models for natural images may only provide limited improvements in accuracy for substantial computational cost.

4.3 Applying HVS Models to Natural Scenes

The previous section outlined some areas where vision models can be improved when being used in compressed image quality evaluation tasks. This section examines these points in detail and, by considering the structure of natural images, suggests appropriate ways in which these factors should be modeled.

As discussed in Sections 2.2.1 and 4.2.1, the contrast in natural scenes should be measured locally and separately within a number of frequency bands. This is supported in a study by Tolhurst and Tadmor [244], which examined human observer’s ability to discriminate changes in the slopes of amplitude spectra in natural scenes. A contrast model based on Peli’s LBC algorithm was shown to accurately predict human performance, whereas other simpler contrast metrics did not. The exact bandwidth of the contrast metric was found not to be critical, as long as it was in the range 0.6–1.5 octaves. Other studies have also shown the improved model accuracy achieved by using the LBC algorithm rather than simpler metrics, when complex images are used [197, 198].

The choice of an appropriate CSF is also important for accurate vision modeling
in natural scenes. Because typical scenes are very complex, it is difficult to find simple, artificial stimuli which can be used for the CSF test to predict contrast sensitivity accurately over the whole scene. However, since the general nature of common coding errors is known (and our CSF target stimuli should resemble coding errors), some general guidelines for stimuli used in the CSF test can be identified and are listed below.

- Stimuli should be of limited spatial extent rather than full-field. Gabor patches are therefore considered more appropriate than full-field sine-wave gratings.

- For still image metrics, the temporal extent of the stimuli should not be too short, since under subjective testing conditions subjects can view the scene for a considerable length of time.

- Temporal presentation type (gradual or abrupt) should reflect the way in which the original and coded images are shown during subjective testing. If the subject can freely alternate between original and coded images, then abrupt is more appropriate. Otherwise, gradual would be more suitable.

- The mean luminance must be varied to cover the extent of luminance values in a typical natural image. Particularly at low luminance levels (< 10 cd/m²), the CSF is known to drop quickly (as it enters the DeVries-Rose region). This effect is also known to be frequency dependent: the DeVries region persists into higher luminance levels for larger spatial frequencies.

If visual stimuli such as these are used to test contrast sensitivity, the resultant CSF curves have a relatively low peak, and are more low-pass rather than band-pass in shape. This is in general agreement with the reduction in CSF at low- and mid- spatial frequencies found by Webster and Miyahara [274] in their contrast sensitivity tests, using observers adapted to natural stimuli.
Spatial masking is another effect which is strongly influenced by the visual stimuli which are used. Spatial masking studies in the literature suggest that the spatial, spatial frequency, and orientation extents of spatial masking can be strongly influenced by the stimulus (see Section 2.2.3 for a review). The spatial frequency extent of masking has been reported to vary significantly: from within channel only [140], to across orientations [88, 90], and across spatial frequencies [150, 228]. In Section 4.1.3 a review of papers which tested a number of different masking models for natural scenes was given. In all cases, the best prediction of subjective performance was achieved by a model which used wideband masking (i.e. masking across different orientations, and sometimes also across spatial frequencies) [3, 74, 80]. This therefore appears to be the most appropriate masking technique for natural images. As well as the extent of masking, its strength must also be considered. Stimulus uncertainty has a strong influence on the strength of the masking effect. Textured regions and areas of high uncertainty induce higher masking (TE gradient ≈ 1.0) than areas of lower uncertainty of the same contrast (TE gradient ≈ 0.7), such as lines and edges. This is particularly significant in natural images, which typically consist of textured regions bordered by edges.

A model of early visual processes is suitable when image fidelity is being assessed. However when the suprathreshold quality of a coded picture is being evaluated, higher level perceptual and cognitive factors also need to be considered. Section 4.2.2 discussed the importance of eye movements and visual attention in determining the overall quality of a compressed scene. In Chapter 6 a model of visual attention and eye movements is proposed, and in Chapter 7 this model is incorporated with the early vision model of Chapter 5, into a new HVS-based metric for image quality assessment.
Chapter 5

A New Early Vision Model
Tuned for Natural Images

Models of the human visual system have been identified as a promising technique for assessing the fidelity and quality of compressed natural pictures. A review of previous quality metrics based on early vision models was given in Chapter 4. In general, it was found that while these techniques could predict the visibility thresholds for simple visual stimuli, their accuracy decreased when complex natural images with suprathreshold distortions were used. Reasons for this change in performance, and suggestions for the application of vision models to natural scenes, were discussed in the latter part of Chapter 4.

In this chapter a new early vision model for assessing both the fidelity and quality of compressed natural images is presented. The model is based on the operation of neurons in the primary visual cortex. Its basic structure is similar to that of previous multi-channel early vision models. However, the components of the model have been tuned specifically for application with natural images. This is important, since the adaptive and non-linear nature of the HVS means that models which are based on data using simple, artificial visual stimuli can perform
poorly when applied to natural scenes. Although the model presented here can be used independently to measure picture quality, in Chapter 7 it is combined with the attention model of Chapter 6 to produce a more robust quality metric.

In Section 5.1 the operation of the model is discussed. This includes a detailed description of the individual model components, along with alternative implementations which may also be suitable. To assess the accuracy of the quality metric, a database of subjective test data is required. Therefore, subjective quality testing was performed, and details of these subjective tests are given in Section 5.2. In Section 5.3, the accuracy of the proposed early vision model is assessed by comparing its predictions with subjective test data. An analysis of the influence that specific model components have on overall accuracy is also given.

5.1 Model Description

A block diagram showing the basic operation of the early vision model is shown in Figure 5.1. It is an extension of the model described in [193]. A brief overview of the model’s operation will now be given. A detailed discussion of the methods used to implement individual model components is provided later in this section.

The luminance values of both the original and coded images are input into the model, along with display and viewing parameters for model calibration. Both input images are converted to band-limited contrast using Peli’s LBC algorithm [195] with a modification suggested in [153]. As discussed in Sections 2.2.1 and 4.2.1, this method of evaluating contrast is well suited for representing contrast in natural images. The LBC algorithm decomposes the images into 1-octave bandwidth channels. These channels can be further decomposed into orientation selective channels by using fan filters [44, 269].

CSF data obtained from psychophysical tests using naturalistic visual stimuli [155,
172, 215, 216] is used to determine the frequency-dependent threshold of detection. This threshold can be raised by spatial masking, which is modeled as a threshold elevation process. Thus detection thresholds are raised in areas of high contrast, with textured regions having a higher rate of threshold increase than edge regions. This new threshold is used to determine whether the difference between the original and coded images is visible. This is done (in each channel and at each location) by subtracting the original and coded LBC signals, and dividing this difference by the detection threshold for that band and location.

By this stage, an indication of whether the distortion was visible has been determined for each band and each location in the image. To produce a single fidelity map indicating the location of visible distortions, Minkowski summation is performed across channels. This produces a map which is the same size as the original image, and indicates whether a distortion between the original and coded image was visible at each location. This fidelity map is termed a *Perceptual Distortion Map* (PDM). The magnitude of the map at a particular location
indicates the number of JNDs that the distortion is above detection threshold; this therefore gives an indication of the perceptual severity of the distortion. To determine the correlation of the metric with subjective MOS data, a single number must be produced. This is accomplished by a further Minkowski summation across space. This single number is termed a *Perceptual Quality Rating* (PQR), which is intended to correlate well with subjective MOS. It can be scaled to the range 1–5 if desired.

The details of the individual components of the model are now discussed in detail.

### 5.1.1 Inputs to the New Model

The luminance values (as displayed on the viewing device) of both the original and coded images are required as inputs to the model. If the image is displayed on a monitor, a conversion of the image from greylevels to luminance is therefore required. The luminance response of the monitor (typically referred to as monitor gamma) needs to be determined, and an appropriate response function or look-up table used for conversion. Other parameters related to the viewing conditions need to be considered to calibrate the model. These include the viewing distance, pixel spacing, image dimensions, and ambient luminance.

### 5.1.2 Channel Decomposition and Band-limited Contrast

Peli’s LBC algorithm [195] is used to decompose the images into multiple channels, and to represent the images as local band-limited contrasts. The spatial frequency decomposition is implemented using cosine log filters (Figure 5.2(a)). These filters are 1-octave bandwidth, and the original image can be fully reconstructed by simple addition. No sub-sampling of the images at lower frequencies was performed, although this can be done for computational efficiency with
Figure 5.2: Structure of filterbanks. (a) Spatial frequency response of filterbank with 1-octave bandwidth cosine log filters (adapted from [195, Figure 8]), and (b) orientation response of fan filters with 30 deg bandwidth (adapted from [44, Figure 14.13]).

minimal loss of accuracy. The Fourier transform $F(r, \theta)$ of the image $f(x, y)$ is calculated, and filtering is then performed using a cosine log filter. This filter, with 1-octave bandwidth and centred at frequency $2^{k-1} \ c/\text{deg}$, is expressed as

$$G_k(r) = \frac{1}{2} \{1 + \cos(\pi \log_2 r - \pi (k - 1))\}, \quad (5.1)$$

where $r$ is the radial spatial frequency. A band-limited version of the image $A(r, \theta)$ is then given by

$$A_k(r, \theta) = F(r, \theta) G_k(r). \quad (5.2)$$

Orientation sensitivity of the channels can be implemented via fan filtering [44, 269] (see Figure 5.2(b)). The fan filter for orientation $\theta$ is given by

$$fan_\ell(\theta) = \begin{cases} \frac{1}{2} \{1 + \cos\left[\frac{\pi (\theta - \theta_c(l))}{\theta_{tw}}\right]\} & \text{if } |\theta - \theta_c(l)| < \theta_{tw}, \\ 0.0 & \text{otherwise,} \end{cases} \quad (5.3)$$
where $\theta_{tw}$ represents the angular transition width and $\theta_c(l)$ is the orientation of the peak fan filter $l$ given by

$$
\theta_c(l) = (l - 1)\theta_{tw} - \pi/2. 
$$

(5.4)

The transition width is set equal to the angular spacing between adjacent fan filters, to ensure that the filtering is fully reversible. The number of fan filters used is typically 4 or 6, which results in orientation bandwidths of 45 deg and 30 deg respectively. Such values agree with the mean bandwidths found in cortical neurons of the monkey and cat [53, p. 266]. As was discussed in Section 4.2.3, there is significant evidence to suggest that the inclusion of orientation sensitivity in the filtering has only a limited influence on the accuracy of quality metrics for natural scenes. Orientation selectivity however introduces a significant increase in computational complexity. For these reasons, the results presented in the remainder of this thesis were obtained using a model which does not include orientation sensitivity. This does not affect the validity of the research findings in other areas of this thesis as the choice of orientation sensitivity is independent of these factors. However, if orientation selectivity is required, then it can easily be implemented using the fan filtering technique described above.

Each spatial frequency band $A_k(r, \theta)$ is transformed back into the spatial domain via an inverse Fourier transform to give $a_k(x, y)$. For the results presented in this thesis, six frequency bands are used (i.e. $k=1,2,..,6$). The local band-limited contrast $c_k(x, y)$ is then calculated as

$$
c_k(x, y) = \frac{a_k(x, y)}{l_k(x, y)}, 
$$

(5.5)

where $l_k$ is the local luminance mean given by

$$
l_k(x, y) = a_0(x, y) + \sum_{i=1}^{k-2} a_i(x, y), 
$$

(5.6)
where $a_0$ is the residual low frequency image. The summing of images up to two bands below the current band (rather than up to the next lowest band) was suggested by Lubin [153] and has been shown to provide a more accurate contrast measure for natural scenes.

### 5.1.3 Contrast Sensitivity

The properties of the visual stimuli used in psychophysical tests have a strong influence on the shape and peak sensitivity of the spatial CSF (see Section 2.2.2 for a review). It is therefore necessary to choose CSF data obtained using visual stimuli which are appropriate for the task that the vision model is designed for. In Section 4.3, such visual stimuli were identified for the task of fidelity and quality assessment of compressed natural images.

Rovamo et al. have comprehensively examined the influence that grating area, exposure time, and light level (retinal illuminance) have on the spatial CSF [155, 172, 215, 216]. Their CSF data is used to establish the base threshold of detection for the early vision model. The spatial extent of the grating patch used was small (1 cycle), and the temporal presentation time was 1.0 seconds. These values are appropriate for compression errors in complex natural images. The influence of light level also needs to be considered. The CSF is stable and follows Weber's law at higher light levels (typically $>10$ cd/m$^2$), but at lower levels it reduces in proportion to the square root of the luminance (DeVries-Rose law). This has been modeled as

$$CSF(l) = \begin{cases} 
    CSF_{base} & l \geq l_{th}, \\
    \sqrt{l/l_{th}} CSF_{base} & l < l_{th},
\end{cases} \quad (5.7)$$

where $l$ is the local luminance, $CSF_{base}$ is the base threshold for higher luminance values (from Rovamo et al. [155, 172, 215, 216]), and $l_{th}$ is the cut-off
Figure 5.3: Typical threshold elevation caused by spatial masking. Log-log slopes of 0.7 and 1.0 are shown.

Luminance below which DeVries-Rose law holds. This cut-off value does have some dependence on spatial frequency (see Section 2.2.2.1), but in the current implementation a value of 10 cd/m$^2$ is used for all spatial frequencies, which provides a good approximation. The CSF data of Rovamo et al. also correlates well with that of Peli et al. [199], which was taken using Gabor patch stimuli of limited spatial extent, and both gradual and abrupt temporal presentation.

5.1.4 Spatial Masking

A comprehensive discussion of spatial masking effects was given in Section 2.2.3. The main factors which influence the amount of masking are the contrast of the background and the uncertainty created by the background. In the early vision model presented here, spatial masking is modeled as a threshold elevation process (see Figure 5.3). As discussed in Sections 2.2.3 and 4.3, simple and predictable (i.e. low uncertainty) stimuli exhibit a TE curve with a log-log slope of around 0.7, while complex (high uncertainty) stimuli have a log-log slope of around 1.0.
This explains why we can tolerate greater error in textured areas than along edges of the same contrast. We model the effects of spatial masking by locally classifying the image into flat, edge, and texture regions. This is performed using a Sobel edge detector locally at four orientations for each point in the image. If the local contrast within a region is below that required to induce masking (i.e. contrast \(< C_{th0}\)), the point is classified as flat. Otherwise, if one of the orientations is dominant, the point is classified as belonging to an edge, else it is classified as belonging to a texture. An orientation was considered dominant if contained more than 40% of the edge energy. Alternative strategies for classifying edges and textures may also be used (e.g. [98, 235]), but it is not clear that any particular method provides superior results. Masking is then implemented by threshold elevation as

\[
C_{th_k} = \begin{cases} 
C_{th0_k} & \text{if } C_{m_k} < C_{th0_k}, \\
C_{th0_k} \left( \frac{C_{m_k}}{C_{th0_k}} \right)^\varepsilon & \text{otherwise},
\end{cases}
\]

(5.8)

where \(C_{th_k}\) is the detection threshold in the presence of a masker for frequency band \(k\), \(C_{th0_k}\) is the base threshold for frequency band \(k\) (from CSF), \(C_{m_k}\) is the contrast of the masker for frequency band \(k\), \(\varepsilon = 0.7\) for edge areas, and \(\varepsilon = 1.0\) for textured areas. This maintains the base detection threshold in flat areas, increases the threshold with contrast gradually along edges, and increases threshold more rapidly in textured regions. The “dipper” effect (see Section 2.2.3 and Figure 2.15), which occurs near threshold when the masker and stimulus are of the same orientation, is not modeled by the threshold elevation function. However this effect is usually minor, particularly in natural scenes (see Section 4.2.3).

The extent of masking in terms of spatial frequency and orientation must also be considered. As discussed in Section 4.3, masking models which incorporate wideband masking (i.e. masking across different orientations, and perhaps also across spatial frequencies) have achieved significantly better results in natural scenes than masking models which consider only within-channel masking. In
the model presented here, orientation selectivity is not implemented, so across-orientation-channel masking occurs implicitly. If the model were implemented with orientation selective channels, then across-orientation-channel masking could be implemented quite easily by including inhibitory effects from any channel which has the same spatial frequency, but different orientation to the channel being tested. Masking across spatial frequencies was not tested in this model, but it could easily be implemented by including inhibitory effects for channels which coincide spatially, but have different spatial frequencies.

Masking will therefore change the threshold of detection of the error in the scene, depending on the local scene content. This new threshold is then used to determine, for each frequency band and at each location, whether the difference between the original and coded images is visible:

\[ V_k(x, y) = \frac{c_k(x, y) - c'_k(x, y)}{C_{th_k}(x, y)}, \]

where \( V_k \) represents the error visibility in each band, \( c_k \) is the original image contrast, and \( c'_k(x, y) \) the coded image contrast. Therefore, when \( V_k(x, y) = 1.0 \), the distortion in frequency band \( k \) at location \( (x, y) \) will be at the threshold of detection.

### 5.1.5 Summation

Following the division of the error signal by the detection threshold for each channel and location in the image, a multi-resolution representation of the visible distortions is produced. To produce a single map showing the visibility of distortions at each location in the scene, it is therefore necessary to perform a summation across channels. Minkowski summation has been shown to be an efficient means for combining the responses of the various channels in the visual cortex [185]. In the current model, Minkowski summation is employed across the
different channels as

\[
PDM(x, y) = \left( \sum_{k} V_k(x, y)^{\beta_1} \right)^{1/\beta_1},
\]

where the PDM represents the visibility of the error at each point, and \( \beta_1 \) is the probability summation exponent with a value of 4.0 [268]. This PDM map identifies the location and magnitude of visible distortions in the scene, and essentially represents the *fidelity* of the image.

In many instances a single number relating to overall picture quality is required. This also allows a quantitative assessment of the vision model, since the correlation of this quality figure with subjective MOS data can be determined. Minkowski summation over all \( N \) pixels has been shown to be an efficient way of converting the PDMs to a single number [48], in effect performing a conversion from fidelity to quality. This is performed using

\[
PQR = \left( \frac{1}{N} \sum_{x, y} PDM(x, y)^{\beta_2} \right)^{1/\beta_2},
\]

where PQR represents the Perceptual Quality Rating. The value of \( \beta_2 \) is used to control the output. A value of \( \beta_2 = 1.0 \) would result in a quality proportional to the average of errors over all pixels, while for high values of \( \beta_2 \), the quality becomes proportional to the maximum PDM value. Following correlation with subjective test data, a value of \( \beta_2 = 3.0 \) was found to give most accurate results. This is in agreement with the subjective data of Hamberg and de Ridder [103]. However, overall correlation did not change significantly if \( \beta_2 \) was increased to 4.0.

Finally, to produce a result on a scale from 1–5 and to enable easy comparison with MOS data, the summed error can be scaled as follows:

\[
PQR_{1-5} = \frac{5}{1 + p \cdot PQR}
\]

\[\text{114}\]
where $\text{PQR}_{1-5}$ represents the PQR scaled to the range $[1.0-5.0]$ and $p$ is a scaling constant. As a result of subjective testing, a value of $p = 0.8$ was found to give best correlation with subjective opinion.

5.2  Subjective Image Quality Testing for Model Validation

A problem with perceptual vision models, particularly when used with supra-threshold visual stimuli, is that it is difficult to assess their accuracy quantitatively. This issue was discussed in Section 4.1.3. Currently the best way to assess the accuracy of a HVS-based fidelity or quality metric is to first decompose its multi-dimensional output into a single number (e.g. convert the PDM to a PQR via summation), and then measure the correlation between this figure and subjective MOS data. It is therefore important to have a substantial database of subjective ratings data which has been gathered under controlled conditions. As no such database is publicly available, subjective testing was required.

The subjective quality evaluation tests were performed in accordance with the regulations specified in Rec. ITU-R BT.500-6. The tests were taken by 18 subjects, exceeding the minimum 15 which is suggested in the Rec. 500 standard. Of these, 16 were unfamiliar with digital image and video compression, while 2 were postgraduate students who were familiar with typical compression artifacts, but picture quality assessment was not part of their normal work.

5.2.1  Viewing Conditions

The viewing room was windowless and draped on all four sides by grey curtains. It was lit by a single incandescent lamp, which was powered by a variable DC
supply and enclosed in a box on the ceiling of the room. An opaque mask was used to ensure that no light shone directly on the monitor itself, but it did shine on the background behind the monitor. The luminance of the background curtains was measured using a Hagner Universal photometer, and was adjusted until it was in the range specified in Rec. 500. The images were displayed on a Sony PVM-2130QM 21-inch monitor, which was controlled by a Silicon Graphics O2 workstation. The front of the monitor was placed flush with the background curtains. The display brightness and contrast were adjusted according to ITU-R BT.814 and ITU-R BT.815, and the luminance response of the monitor over the full range of greylevels was recorded.

The subjects viewed the monitor from a distance of five screen heights. Up to three subjects could be tested at the same time.

5.2.2 Viewing Material

Four different scenes were used in the test, and they are shown in Figures A.1(a), A.2(a), A.8(a), and A.11(a). These particular scenes were chosen as they are dissimilar from one another, and give a reasonable representation of the types of images encountered in practical situations. Both JPEG [120] and wavelet [79] coding were used to introduce distortions into the pictures. A wide range of bit-rates were tested (0.15–1.65 bit/pixel), which contained subjective image qualities ranging from ‘bad’ to ‘excellent’ on the Rec. 500 quality scale (see Figure 5.4). In order to provide a more challenging test for the quality metric, composite images with spatially varying quality were also used. These were obtained by manually selecting rectangular ROIs in each scene, cutting these regions from a high bit-rate picture, and pasting them onto a low bit-rate version of the picture. No attempt was made to smooth the pasted region into its surrounds. The resultant image has high quality in the ROIs, and lower quality in the peripheral areas. Between one and three rectangular ROIs were obtained in each in the four scenes
Figure 5.4: Sample picture quality rating form used in the DSCQS tests.

used for these subjective tests. The final test set consisted of 16 JPEG coded images, 16 wavelet coded images, and 12 composite images coded with variable quality, which produced a final set of 44 test images.

5.2.3 Testing Procedure

Testing was performed in compliance with the Rec. 500 DSCQS method. A timing diagram for the presentation of the images is shown in Figure 5.5. The original and coded scenes were shown in successive pairs; however the order in which the original and coded scenes were presented was pseudo-random and not known to the subjects. The ordering of the 44 test scenes was also pseudo-random and was changed for different observers. The original and coded scenes were displayed for around five seconds each, and this process was repeated three times. Subjects
were then given time to vote on the quality of each of the images (A and B), using the continuous quality scale shown in Figure 5.4. A training set of six images was used to familiarise the subjects with the testing procedure. However the results of this training set were not used in the final analysis. The duration of a complete test session was around 25 minutes.

### 5.2.4 Results of DSCQS Tests

The raw subjective data was analysed and processed using the methods specified in Rec. 500-6. The ratings for the individual subjects are tabulated in Appendix B. These quality ratings were averaged across all subjects, and the results for the four images are shown in Figures 5.6 and 5.7. As the variable quality scenes consisted of a composite of two coded pictures, their bit-rates had to be approximated. This was calculated by adding the proportion of bits (with respect to area occupied) of the low and high quality scenes from which the variable quality scene was composed. As can be seen from Figures 5.6 and 5.7, variable quality scenes were always given a higher subjective rating than uniformly coded scenes of the same bit-rate. The extent of this effect varied significantly between the scenes (from 0.10–1.40 MOS ratings). Another general trend was that at high bit-rates, JPEG coded scenes typically had a higher subjective quality than wavelet
Figure 5.6: Subjective MOS data averaged across all subjects for the first two test images (a) announcer and (b) beach. Error bars represent the 95% confidence interval of the mean MOS rating.
Figure 5.7: Subjective MOS data averaged across all subjects for the final two test images (a) lighthouse, and (b) splash. Error bars represent the 95% confidence interval of the mean MOS rating.
coded scenes of the same bit-rate. However, as the bit-rates were decreased the
difference in MOS between JPEG and wavelet decreased, and at low bit-rates the
wavelet coder typically gave better subjective quality than JPEG. This agrees
with the subjective data obtained by Avadhanam and Algazi [12].

A strong difference in subjective ratings for individual images can also be iden-
tified. Simpler scenes such as splash and announcer degrade in quality more
gracefully than the complex beach and lighthouse scenes. This demonstrates that
bit-rate alone is a poor indicator of picture quality.

5.3 Performance of the Early Vision Model

The accuracy of the early vision model, and the significance of different model
configurations, have been tested using two general strategies. Firstly, the PDMs
produced by the model have been visually inspected, to ensure that the location
and magnitude of predicted distortions correspond qualitatively with subjective
opinion. Secondly, the correlation of the model’s PQR predictions with the sub-
jective MOS data described in the previous section has been determined and
analysed.

The early vision model has been tested using a wide range of different scenes, bit-
rates, and coders. A typical example of its output, in comparison to the squared
error, can be seen in Figure 5.8. The fur in the mandrill image is very strongly
textured. Significant distortion can therefore be tolerated in these areas without
them being visible, since the strong textures will mask distortions. However
the areas around the nose and cheeks of the mandrill are relatively smooth, so
the threshold of detection in these areas is much lower. The coded image in
Figure 5.8(b) has visible distortions around the nose and cheeks, and also in
the less textured areas in the lower extremities of the image. This corresponds
well with the visible errors detected by the PDM. However, the problems of
Figure 5.8: Fidelity assessment for the image *mandrill*. (a) Original image, (b) JPEG coded at 0.7 bit/pixel, (c) PDM, and (d) squared error. In (c) and (d), lighter regions represent areas of visible distortion.
MSE-based metrics are clearly evident in Figure 5.8(d). Since the squared error doesn’t take into account any masking or luminance non-linearities, it is unable to model our reduced sensitivity to distortions in textured areas, such as in the fur of the mandrill. MSE-based metrics are therefore over-sensitive to distortions in textured parts of the scene, which are the areas where most coders introduce the largest numerical errors.

The correlation of the PQR with the subjective test data of Section 5.2 has been determined and is plotted in Figure 5.9. The correlation of the MSE with the subjective MOS has been included for comparison. This plot clearly demonstrates the inadequacy of the MSE technique for measuring picture quality across different scenes and coders. The MSE technique operates monotonically with bit-rate for a particular coder and scene. However, since it does not take into account the structure of the scene or the distortion, results across scenes and coders cannot be reliably compared. The PQR technique does not exhibit such a strong dependency on scene and coder, and is therefore quite robust. It does however underestimate the quality of the variable coded scenes, as it does not take into account the location of errors with respect to the scene.

The raw correlations with subjective MOS achieved by these techniques are shown in Table 5.1. Across all scenes (both uniform and variable coded), PQR achieves a correlation of $|r| = 0.87$ with MOS, while MSE only has a correlation of $|r| = 0.65$. Both techniques perform slightly better when only the uniform coded scenes are considered, and worse when only variable coded scenes are used. This can be

<table>
<thead>
<tr>
<th>Objective Metric</th>
<th>Set of Images</th>
<th>Uniform Coded Only</th>
<th>Variable Coded Only</th>
<th>All scenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>PQR</td>
<td>$</td>
<td>r</td>
<td>= 0.94$</td>
<td>$</td>
</tr>
<tr>
<td>MSE</td>
<td>$</td>
<td>r</td>
<td>= 0.74$</td>
<td>$</td>
</tr>
</tbody>
</table>

**Table 5.1:** Correlation of PQR and MSE with subjective MOS.
Figure 5.9: Correlation of objective quality metrics with subjective MOS. (a) PQR, and (b) MSE. ‘○’ = uniform JPEG; ‘×’ = uniform wavelet; ‘□’ = variable JPEG; ‘◦’ = variable wavelet.
expected, since neither the PQR nor the MSE consider the location of distortions relative to the scene content.

Statistical tests were performed to determine the significance of these correlations. The first test was to determine whether the correlations achieved by PQR and MSE were statistically significant \((H_0 : \rho = 0)\). This was done using a \(t\)-test with \(v = n - 2\) degrees of freedom:

\[
t = \frac{r}{\sqrt{1 - r^2} / n - 2}.
\]

The correlation of \(|r| = 0.87\) achieved by PQR is significant at the \(\alpha = 0.001\) level \((t = 11.44; t_{0.0005, 42} = 3.55)\). Therefore the null hypothesis \(H_0 : \rho = 0\) can be rejected at all significant levels of \(\alpha\). The correlation of \(|r| = 0.65\) achieved by PQR is also significant at the \(\alpha = 0.001\) level \((t = 5.54; t_{0.0005, 42} = 3.55)\). To test whether the improvement in correlation achieved by PQR over MSE is statistically significant, the hypothesis \(H_0 : \rho_1 = \rho_2\) is tested using the Fisher Z independent correlation test:

\[
z = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}},
\]

where \(z\) is critical value of the Normal distribution, and \(Z_1\) and \(Z_2\) are Fisher Z scores given by

\[
Z = \frac{1}{2} \ln \left( \frac{1 + |r|}{1 - |r|} \right).
\]

This produces a value of \(z = 2.52\) for the difference in correlation of PQR and MSE, which is significant at the level \(\alpha = 0.01\) \((z_{0.01} = 2.33)\). Thus the null hypothesis \(H_0 : \rho_1 = \rho_2\) can be rejected at the level \(\alpha = 0.01\), and it can be concluded that the increase in correlation achieved by the PQR metric is statistically significant at this level.
5.3.1 Influence of the Choice of CSF on the Vision Model

To determine the effect that the choice of CSF data has on both the PDM and PQR, the model was tested using both the full-field sine wave grating CSF data of van Nes and Bouman [259] and the patch CSF data of Rovamo [155, 172, 215, 216] (see Figure 2.11). The full-field data of van Nes and Bouman is approximately five times more sensitive than that of Rovamo et al. at medium spatial frequencies, and the peak of the CSF is also at a slightly higher spatial frequency. A visual inspection of the PDMs produced using both types of CSF data clearly indicates that the van Nes and Bouman data is over-sensitive to distortions in complex natural images. However, when the Rovamo et al. data is used, the PDMs of scenes coded at the threshold of detection correlate well with subjective opinion. An example of this can be seen in Figure 5.10. The coded scene in Figure 5.10(b) was one of the images used in the subjective testing described in Section 5.2. Its average MOS was rated as 4.97, which suggests that some distortions in this scene have just started to become visible. The PDM produced using the Rovamo et al. data qualitatively agrees with this. However the PDM produced using the van Nes and Bouman data is over-sensitive to the distortions in the complex scene, since it predicts that errors are well beyond the threshold of visibility in many areas of the scene. A similar over-sensitivity of the van Nes and Bouman data was found for all images which were coded at the level where their subjective MOS just began to decrease below 5.0.

Although the use of full-field CSF data produces an over-sensitive vision model, the overall correlation with MOS did not decrease when the van Nes and Bouman CSF data was used. This suggests that the vision model scales quite linearly when it moves into suprathreshold levels of distortion. Therefore the over-sensitive full-field CSF data can still provide a good correlation with MOS, since it effectively is just a scaling of the model which uses patch CSF data. However, using a full-field CSF is likely to cause problems for the vision model when used with very low and
very high distortion images. For scenes with very low distortion (i.e. below the
threshold of detection), the full-field CSF model may still predict some visible
distortions, and will therefore suggest a reduction in subjective quality when it is
not warranted. A model tuned with patch CSF data would avoid this problem,
since its PDM would be flat (zero) for sub-threshold distortions. At high levels of
distortion, the model using full-field CSF data is likely to encounter non-linearities
sooner than the model using patch CSF data, since it predicts larger JND values.
5.3.2 Influence of Spatial Masking on the Vision Model

The spatial masking technique used in an early vision model will have a significant effect on the overall accuracy achieved by the model. To demonstrate this, the model was implemented with the spatial masking turned off (i.e. the threshold elevation slope was set to zero for all levels of background contrast). The correlation with MOS across all images reduced sharply to $|r| = 0.77$. Visual inspection of the PDMs confirmed that, if no spatial masking was implemented, the vision model predicted too much visible error near strong edges, and particularly in highly textured regions.

The significance of using different threshold elevation slopes for edge and texture regions (0.7 and 1.0 respectively) was also investigated. When the same TE slope of 0.7 was used for both edge and texture regions, the overall correlation with MOS dropped slightly to $|r| = 0.86$. Visual inspection of the PDMs suggested that, when no distinction is made between the spatial masking in edge and textured regions, the model is over-sensitive to distortions in textured regions. This can be seen in a frame from the football sequence in Figure 5.11. This scene has distortions in the grass which are just visible. The textured regions in the grass induce more masking than is predicted by the model with the same TE slope of 0.7 for both texture and edge regions, so the PDM is consequently over-sensitive in these areas. However the model which uses different TE slopes for edge and texture regions (0.7 and 1.0 respectively) correctly estimates the higher masking in these textured areas, while maintaining sensitivity near edges.

5.3.3 Analysis of the New Early Vision Model

The visual inspection of the PDMs produced by the new early vision model indicated that it is a good measure of image fidelity for complex natural images. Across a wide range of scenes, coders, and bit-rates, the model's predictions of
Figure 5.11: Fidelity assessment for the image football. (a) Original image, (b) JPEG coded at 1.0 bit/pixel, (c) PDM using TE slopes of 0.7 and 1.0 for edges and textures respectively, and (d) PDM using the same TE slope of 0.7 for both edges and textures. In (c) and (d), lighter regions represent areas of visible distortion.

both the location and magnitude of visible errors were in strong agreement with subjective appraisal of visible errors. The choice of appropriate CSF data for the model was found to strongly influence the accuracy of the PDM. CSF data from full-field sine wave grating experiments produced an over-sensitive model. However, when CSF data obtained using appropriate stimuli was used, the PDMs accurately predicted the threshold of detection. The choice of an appropriate spatial masking technique for natural images was also found to affect the accuracy of the PDMs. When no distinction was made between spatial masking in edge and
texture regions, the resultant PDMs were over-sensitive to distortions in textured regions. However, when the model introduced higher masking in textured regions, this over-sensitivity was eliminated.

Summation of the visible errors in the PDM to produce a PQR was performed using a simple Minkowski summation. Looking at Figure 5.9(a), the dense congregation of data points at high qualities suggests that this simple approach is quite effective for scenes with low distortion. However, overall correlation with MOS decreased substantially for scenes with medium to high distortion, and some scene and coder dependencies were evident. In particular, the Minkowski summation had problems with scenes coded with variable quality, typically under-estimating their subjective quality. The problems caused by using such a simple summation can be expected, since it treats errors of equal magnitude on the PDM equally, irrespective of their relative location or structure in the scene (see Section 4.2.2). Such an approach fails to consider any higher level or cognitive factors, and will inherently have problems predicting overall image quality, particularly in scenes with suprathreshold distortions.

In summary, the new early vision model presented here provides a good estimate of the fidelity of compressed natural images. Some improvements in accuracy may be obtainable by increasing the model’s complexity or through further tuning of model parameters. However the improvements in quality estimation achieved by doing this are likely to be minimal. An area which so far has received little attention, but which promises to significantly improve quality assessment accuracy, is the analysis and summation of the JND maps to produce an overall estimate of picture quality. Simple techniques such as Minkowski summation fail to take into account higher level and cognitive processes. Consequently, they suffer from inaccuracy when suprathreshold and complex distortions are present. By including higher level processes in the vision model, significant improvements in model accuracy are likely to be attainable. This issue is addressed in the next two
5.3 Performance of the Early Vision Model

chapters. A model of visual attention is presented in Chapter 6, which can automatically detect ROIs in a scene. This attention model is used to weight the outputs of the early vision model prior to summation, in a new quality metric proposed in Chapter 7.
Chapter 6

Identifying Important Regions in a Scene

Models of the early visual system which are tuned appropriately can provide accurate predictions of the location of visible distortions in compressed natural images. To produce an estimate of subjective quality from these fidelity maps, current state-of-the-art quality metrics perform a simple summation of all visible errors. This fails to take into account any higher level or cognitive factors which are known to occur during the subjective assessment of picture quality.

The influence that a distortion has on overall picture quality is known to be strongly influenced by its location with respect to the scene content. The variable-resolution nature of the HVS means that high acuity is only available in the fovea, which has a diameter of around 2 deg. Knowledge of a scene is obtained through regular eye movements, to reposition the area under foveal view. Early vision models assume an “infinite fovea”; in other words, the scene is processed under the assumption that all areas could be viewed by the high acuity fovea. However, as detailed in Section 2.3.1.1, studies of eye movements indicate that viewers do not foveate all areas in a scene equally. Instead, a few areas are identified as
ROIs by our visual attention processes, and viewers tend to repeatedly return to these ROIs rather than other areas which have not yet been foveated. The fidelity of the picture in these ROIs is known to have the strongest influence on overall picture quality [65, 96, 136, 252].

As was detailed in Section 2.3, a great deal has been learned about the operation of eye movements and visual attention processes. A number of bottom-up and top-down factors which influence visual attention have been identified (see Section 2.3.4). This knowledge of human visual attention and eye movements, coupled with the selective and correlated eye movement patterns of subjects when viewing natural scenes, provides the framework for the development of computational models of human visual attention. These models aim to detect the important areas in a scene (i.e. the areas where viewers are likely to focus their attention) in an unsupervised manner.

This chapter presents an automatic way to predict where ROIs are likely to be located in a natural scene, using the properties of human attention and eye movements. The model processes the scene to produce an Importance Map, which gives a weighting of the relative likelihood that an area will attract a subject’s attention, for each region in the scene. In Chapter 7 IMs are used to weight the output of an early vision model. This allows the location of distortions to be taken into account by the quality metric, and the resultant improved correlation with subjective opinion demonstrates the utility of this approach. Although the primary purpose of the IM technique in this research was for use in a quality metric, many other applications can also benefit from IMs. In Chapter 8 the IM algorithm is used to control quantisation in an MPEG encoder, and improved subjective picture quality is demonstrated. The conclusion discusses various other applications which could profit from the use of IMs.

The organisation of this chapter is as follows. A review of previous computational models of visual attention and object saliency is given in Section 6.1. The details
of the IM algorithm for still images are provided in Section 6.2, along with results of the algorithm when it is run on typical natural scenes. A validation of the accuracy of the IM algorithm has been undertaken and the results are outlined in Section 6.3. This was performed by correlating the IM’s predictions with subject’s eye movement data (obtained using an eye tracker), for a variety of natural scenes. In Section 6.4, the IM algorithm is extended to video, by taking into account the influence of object motion.

6.1 Previous Computational Models of Visual Attention and Saliency

A steadily increasing number of models of visual attention and object saliency in a scene have been proposed since the mid-1980s. Some have been aimed at simple, artificial scenes (like those used in visual search experiments), while others have been used for natural images. The models can be broken into two general categories: multi-resolution-based and region-based. Both of these techniques are founded on a similar principle, which is to assess the attentional impact of a number of different features (also referred to as factors) and then combine them to produce an overall attention or saliency map. However, the mechanisms by which these calculations are performed are different. Multi-resolution approaches decompose the scene into spatial frequency and/or orientation selective channels, in a manner similar to the processing performed in the primary visual cortex. Attentional features can then be calculated from the multi-resolution decomposition. On the other hand, region-based techniques first segment the scene into homogeneous regions or objects. Attentional factors can then be calculated for each region in the scene. This approach is supported by evidence that a segmentation of a scene into objects occurs at quite an early stage of visual processing [163]. Further support for region-based techniques is obtained from attention studies
which show that subjects are attracted to *objects* rather than *locations* (see Section 2.3.4). Mounting evidence indicates that our pre-attentive vision segments a scene into objects (see [247, 272, 284] for a discussion). Attention to a target is diminished when it is difficult to produce an easy segmentation of the object from the scene.

### 6.1.1 Multi-resolution-based Attention Models

The first computational model of attention was proposed by Koch and Ullman [135]. Their model consists of several stages. First, “feature maps” are calculated for a number of bottom-up features known to influence visual attention: contrast, colour, motion, and orientation. This is performed following a multi-scale decomposition of the scene. Locations which differ significantly from their surrounds with respect to these features are allocated a high conspicuity on the feature maps. The individual feature maps are then combined to produce a “saliency map.” Although Koch and Ullman suggest that the individual feature maps should be weighted differently when calculating the saliency map, no indication is given of how this should be performed. The model then uses the saliency map to predict the temporal order in which subjects would fixate on locations in a scene. The first fixation region is the area of highest salience in the saliency map. To change this fixation location temporally, an IOR model with a suppression duration of 500 msec is used to temporarily inhibit the return of fixation to the same location. Simple Gestalt rules of proximity and similarity preference are also included in the temporal response model. Although the approach used to calculate the saliency map seems valid, the idea of being able to extend this to predict the temporal sequencing of attentional shifts while viewing complex natural scenes has fundamental difficulties. This is because it is known that the temporal order in which objects are fixated varies significantly across subjects (see Section 2.3.1.1).
Since their original saliency model, Koch and his colleagues have proposed a number of extensions to their technique, and provided a clearer description of implementational details [121, 179, 180]. In [179], the feature maps are all weighted equally when being combined to produce the saliency map, except for motion which has an influence five times larger than the other features. In a recent paper [121], examples of the operation of the technique on natural images have been provided, although no validation with subject’s recorded eye movements were performed. Only three features (contrast, colour, and orientation) are used in the latest versions of the model.

Wolfe’s Guided Search models enable the inclusion of top-down as well as bottom-up factors [281, 283, 284, 285]. They were originally designed for the purpose of modeling human response in visual search tasks, and hence have only been used so far with artificial stimuli of low to moderate complexity. Guided Search 2.0 [281] involves the generation of feature maps, following a decomposition into featuresensitive channels. Areas which stand out from their surrounds with respect to a particular feature, and areas with large feature values (“strong” features) are weighted highly in the feature map. In the simulation presented in their paper, only colour and orientation are used as features. However, other bottom-up features could also be included. The model also allows top-down influence, by enabling a higher weighting to be given to features or sub-features if it is known a priori that a subject is attracted to or searching for a particular kind of stimulus. For example, areas which are blue in colour will be weighted higher if it is known that a subject is searching for blue objects. The model has been shown to provide a good prediction of search performance in visual search tasks for artificial stimuli. However, additional features may need to be included if the model is to be used for complex natural scenes.

A model which operates in a similar manner to Guided Search has been proposed by Ahmad [1, 2]. It too combines colour and orientation feature maps, which can
be controlled by top-down influence. This model was also designed for artificial stimuli; consequently no results are reported for natural images.

A simple multi-scale attention model was adopted by Zabrodsky and Peleg [293, 294] for their attention-based video coding representation. A Laplacian pyramid is used to extract high contrast edges, and a temporal pyramid (implemented as a difference of Laplacian) is used to determine regions undergoing motion. The single area which has the highest spatial contrast and motion is identified as the most salient area. A simple IOR function is included to ensure that the region of highest salience changes with time. Some examples of this model’s operation are provided for relatively simple scenes. However, this model is likely to encounter problems in more complex scenes, since it only considers two factors which influence attention.

Another multi-scale attention model was proposed by Milanese [168]. This model produces feature maps for colour, intensity, orientation, and edge magnitude, and combines them to produce an overall saliency map. Provision is also made for the inclusion of motion and top-down control, if such information is available. The model has been designed primarily for machine vision with simple input scenes. The results which were reported for these simple scenes show reasonable quantitative correlation with objects in the scene, and the model shows promise for machine vision applications. However, no correlation with subjective data has been presented, and only limited results have been demonstrated with complex natural images.

Other multi-resolution models have been proposed which rely on the top-down control of bottom-up feature maps [210, 211, 250, 251]. These models report a good correlation with human performance on search tasks when the target is known a priori. However, in the general case, little is known about the content of the scene and about the context with which it was being viewed. Therefore such top-down approaches in which the search target is known a priori are not
suitable for modeling attention in arbitrary natural scenes.

6.1.2 Region-based Attention Models

More recently, a number of region-based models of visual attention have been proposed. Like multi-resolution-based models, this approach is based on the combination of a number of feature maps into an overall attention or saliency map. However, the way individual feature maps are calculated is different. The scene is first segmented into homogeneous regions, and feature maps are then generated for each region in the scene. As mentioned earlier in this section, there are various reasons why a region-based approach is suitable for the prediction of ROIs in a scene.

The first region-based attention model was proposed by Maeder et al. [157, 158]. Following a segmentation of the scene, various factors known to influence visual attention are calculated for each region of the scene. In the example given in their papers, the factors used were edge strength, texture energy, and contrast; however other factors could also be included. These intermediate feature maps are normalised [159], and linearly combined with equal weighting to produce an overall Importance Map. This map was used to control a variable-resolution JPEG encoder, and results showed an increase in subjective picture quality, compared to a conventional JPEG encoder at the same bit-rate.

Marichal et al. [55, 161] developed a region-based attention model for video which uses fuzzy logic rules to control the weightings inside each feature map. Segmentation is first performed using a watershed algorithm. Each of the regions is then classified with respect to five “criteria” affecting attention:

- **Image border criterion**: regions near the border of the scene have a lower weighting than central regions,
- **Face texture criterion**: regions with a colour similar to human skin are weighted highly,

- **Motion and contrast criterion**: highest weighting is given to high contrast regions undergoing medium motion,

- **Interactive criterion**: this is a top-down factor, which gives highest weighting to regions which appear in a user-defined rectangular area of the scene,

- **Continuity criterion**: results from previous frame are taken into account to provide some temporal stability.

An empirical fuzzy weighting scheme is used to develop rules regarding the weighting of these factors within and across the feature maps. However, the values of these parameters are not reported. Some sample results on simple natural scenes (*Akiyo* and *table tennis*) are given. The technique gave quite accurate results for these two scenes. However, these results were influenced strongly by the “interactive criterion.” In a general situation, such top-down information from a viewer would not be available. No results of the model’s performance without this top-down influence have been reported.

Zhao et al. [298, 299] have proposed a region-based attention model for still images. A non-parametric clustering algorithm is used to segment the image into homogeneous regions. Six features are then calculated for each region:

- **Area ratio**: size of the region with respect to the image,

- **Position**: distance from region to the centre of the image,

- **Compactness**: classifies the shape, producing a value of 0.0 for complex shapes, and 1.0 for round shapes,

- **Border connection**: the number of pixels in the region that are also on the boundary of the scene,
- **Region colour**: chrominance values, in the colour space L* a* b*,

- **Outstandingness**: a measure which gives large value to regions which differ from their surrounds with respect to colour, area, and compactness.

To determine the influence that these features have on visual attention, subjective data was collected from 15 subjects. They were presented with both an original scene and a segmented scene for a number of different scenes (20 in total). The subjects were required to classify each region of the scene as being either “very important”, “somewhat important”, or “not important”. This data, was used to train a neural net, in order to determine the influence that the six features have on the overall importance of regions in the scene. The limited results that have been presented for simple natural scenes have indicated that the technique achieves quite good correlation with subjective ROIs. The features which were used in this approach are closely related to some of the factors known to influence attention, which were reviewed in Section 2.3.4. However, since this model is calibrated on subjective data from a limited number of subjects and scenes, it can not be guaranteed to work well for classes of scenes which are outside the range used in the subjective testing. Application of this model across different scene types will require re-calibration on new subjective data.

### 6.2 Importance Map Technique for Still Images

The Importance Map technique which is described here is a region-based attention model. Although it is difficult to provide a quantitative comparison of multi-resolution-based and region-based attention models, the region-based techniques offer several advantages for natural scenes. In general it is easier to include a large number of factors in the model when using a region-based approach. This is because many factors which influence attention are either inherent properties
of regions (e.g. size, shape) or can naturally be associated with whole objects in a scene (e.g. motion, colour, contrast, texture). Recent evidence that attention is associated with objects rather than locations provides further support for a region-based approach (see Section 2.3.4). However there are several disadvantages of a region-based methodology, the most critical one being its reliance on a satisfactory segmentation algorithm. Image segmentation is still an area of active research, and future improvements in segmentation techniques will increase the utility of region-based models. The calibration of attention models is a general problem faced by both multi-resolution-based and region-based models, and is due to a lack of quantitative experimental data. As a result, all current models are calibrated in an empirical fashion.

An overall block diagram showing the operation of the Importance Map algorithm can be seen in Figure 6.1. The original grey-scale image is input (as an array of grey-levels) and is segmented into homogeneous regions. A recursive split and merge technique is used to perform the segmentation. It provides satisfactory results and is computationally inexpensive. However other segmentation algorithms may also be suitable and may provide improved segmentation, at the cost
of computational expense. The split and merge technique uses the local region variance as the split and merge criterion. The splitting process is performed by comparing a block’s variance to the split variance threshold. If it exceeds the threshold, the block is divided into four sub-blocks and the splitting process is invoked recursively on these sub-blocks. This process continues until all blocks either satisfy the split criterion, or their size is smaller than a pre-determined minimum (chosen as 1 pixel in the current implementation). The merging process begins once the splitting has completed. Any adjacent regions will be merged into a single region if the merged region has a variance lower than the merge variance threshold. Split and merge variance thresholds equal to 250 provide good results for most natural scenes, and this value is used as a default. However, the optimal threshold values are scene dependent, so a manual image-specific optimisation of the thresholds will generally provide an improved segmentation. This is the only process in the IM algorithm where manual tuning is suggested; however using the default variance thresholds will in general still provide satisfactory results. To remove the large number of very small regions which may still be present following the split and merge processes, a minimum region size threshold of 16 pixels has been used. All regions that have a size lower than this threshold will be merged with their largest neighbour. In general, a slight over-segmentation will provide better results than an under-segmentation.

The segmented image is then analysed by a number of different factors which are known to influence attention (see Section 2.3.4), resulting in an importance feature map for each of these factors. Five different attentional factors are used in the current implementation. However, the flexible structure of the algorithm facilitates the easy incorporation of additional factors. Thus if a priori information was available about the image or the context of viewing, these factors could easily be added. The different factors which are currently used are listed below.

- **Contrast of region with its surrounds.** As discussed in Section 2.3.4.2, areas
which have a high contrast with their local surrounds are strong visual attractors. The contrast importance $I_{\text{contrast}}$ of a region $R_i$ is calculated as:

$$I_{\text{contrast}}(R_i) = \overline{gl}(R_i) - \overline{gl}(R_{i-\text{neighbours}}),$$  \hspace{1cm} (6.1)

where $\overline{gl}(R_i)$ is the mean grey level of region $R_i$, and $\overline{gl}(R_{i-\text{neighbours}})$ is the mean grey-level of all of the neighbouring regions of $R_i$. Subtraction is used rather than division, since it is assumed that the grey-levels approximate a perceptually linear space. The resulting $I_{\text{contrast}}$ is then scaled so that the maximum contrast importance of a region in an image is 1.0, and the range of $I_{\text{contrast}}$ is $[0.0-1.0]$.

- **Size of region.** Section 2.3.4.4 discussed the influence that object size has on attentional processes. In particular, larger regions were recognised as being strong visual attractors. Importance due to region size is calculated as

$$I_{\text{size}}(R_i) = \min(\frac{A(R_i)}{A_{\max}}, 1.0),$$  \hspace{1cm} (6.2)

where $A(R_i)$ is the area of region $R_i$ in pixels, and $A_{\max}$ is a constant used to prevent excessive weighting being given to very large regions. $A_{\max}$ is equal to 1% of the total image area in the current implementation.

- **Shape of region.** Areas with unusual shape, or areas with a long and thin shape have been identified as attractors of attention (see Section 2.3.4.6). Importance due to region shape is calculated as

$$I_{\text{shape}}(R_i) = \frac{bp(R_i)^{sp}}{A(R_i)},$$  \hspace{1cm} (6.3)

where $bp(R_i)$ is the number of pixels in the region $R_i$ which border with other regions, and $sp$ is a constant. A value of $sp = 1.75$ is used to minimise
the dependency of this calculation on region size. Long and thin regions, and regions with a large number of edges, will have a high shape importance, while for rounder or regular-shaped regions it will be lower. The final shape importance is scaled so that the maximum shape importance of a region in each image is 1.0, and the range of $I_{shape}$ is $[0.0-1.0]$.

- **Location of region.** Several eye movement studies have shown a strong preference for objects located centrally within the scene (see Section 2.3.4.5). Importance due to the location of a region is calculated as

$$I_{location}(R_i) = \frac{centre(R_i)}{A(R_i)},$$

(6.4)

where $centre(R_i)$ is the number of pixels in region $R_i$ which are also in the central 25% of the image (calculated in terms of area). Thus regions contained entirely within the central quarter of an image will have a location importance of 1.0, and regions with no central pixels will have a location importance of 0.0.

- **Foreground / Background region.** In Section 2.3.4.7 several studies were cited which indicated that viewers are more likely to be attracted to objects in the foreground that those in the background. It is difficult to detect foreground and background objects in a still scene, since no motion information is present. However, a general assumption that can be made is that foreground objects will not be located on the border of the scene. Background regions can then be detected by determining the number of pixels which lie on the image border that are contained in each region. Regions with a high number of image border pixels will be classified as belonging to the background and will have a low Background/Foreground importance as
6.2 Importance Map Technique for Still Images

given by

\[ I_{\text{background}}(R_i) = 1.0 - \min\left(\frac{\text{borderpix}(R_i)}{0.5 \times \text{total borderpix}}, 1.0\right), \quad (6.5) \]

where \( \text{borderpix}(R_i) \) is the number of pixels in region \( R_i \) which also border on the image, and \( \text{total borderpix} \) is the total number of image border pixels. Any region which contains greater than half of the total number of image border pixels is likely to be part of the background, and is assigned a Background/Foreground importance of 0.0.

By this stage a feature map has been produced for each of the five factors, indicating the importance of each region in the scene with regards to the factors. Each feature map has been constrained to be contained in the range \([0.0–1.0]\).

The five feature maps now need to be combined to produce an overall IM for the image. As mentioned in Section 2.3.4.13, there is little quantitative data which indicates the relative a priori importance of these different factors. The situation is complicated by the fact that the relative importance of a particular factor is likely to change from one image to the next. In the absence of any firm quantitative evidence, each factor is assigned equal importance in the IM model. However, if it became known that a particular factor had a higher influence, a scaling term could easily be incorporated.

When combining the importance factors, there is once again little quantitative data which would suggest an obvious method for combination. However, it would be preferable to assign a high importance to regions which rank very strongly for any of the factors, since it is known that objects which stand out from their surrounds with respect to a factor are likely to attract attention. A simple averaging of the feature maps is therefore unsuitable. To further enhance the influence of regions which have high salience, the feature maps are raised to an exponent (set to 2.0 in the current implementation), before being summed to produce the final
6.2 Importance Map Technique for Still Images

IM. This is expressed as:

\[ IM(R_i) = \sum_{k=1}^{5} (I_k(R_i))^2, \]

(6.6)

where \( k \) sums through the five importance factors. The final IM is produced by scaling the result so that the region of highest importance in the scene has a value of 1.0.

To ease the abrupt change from an important region to a region of lower importance, and to increase the extent of important regions beyond the region’s edge, it may be suitable to perform post-processing on the Importance Maps. One technique which was found to be useful (see Section 6.3) was to calculate the maximum importance value for every \( n \times n \) block in the map, and assign this value to the whole \( n \times n \) block. Useful values for \( n \) range from 4 to 32 pixels, depending on the size of the image and the purpose of the IM. This process also reduces the impact of inaccuracies in the segmentation.

The IM algorithm has been tested on a wide range of different scene types, and good results have been achieved for both simple and complex scenes. A general observation is that the technique will achieve good results if a satisfactory segmentation of the scene was possible. If the scene was over-segmented, the large number of small regions produced (which do not correspond to whole real-world objects) can result in a degradation in performance. If the scene was under-segmented, distinct objects would be merged into a single region, once again causing some inaccuracy.

Figure 6.2 demonstrates the processes involved in obtaining the IM for the relatively complex \textit{boats} image. As can be seen in Figure 6.2(b), the split and merge segmentation slightly over-segments the image (particularly in the boats on the left of the picture), with 150 separate regions being produced. In addition, some of the very thin boat masts were merged with neighbouring regions during the
Figure 6.2: IM for the *boats* image. (a) Original image; (b) segmented image; feature maps for (c) contrast, (d) size, (e) shape, (f) location, (g) foreground/background; and (h) final IM produced by summing (c)–(g). For (c)–(h), lighter regions represent higher importance.
segmentation. The five importance feature maps are shown in Figures 6.2(c)–(g). The contrast map assigns high importance to regions of high contrast on the side and deck of the boats. The size map allocates high importance to the larger regions, as well as to the background. The shape map assigns high importance to the long, thin masts, and moderate importance to other regions with unusual shapes. The location factor assigns high importance to the centrally-located areas, while the foreground/background factor does a good job of identifying the boats as foreground objects. The five importance factors are combined to produce the final IM in Figure 6.2(h). This map provides a good estimate of the areas which are likely to attract a viewer’s attention in the picture.

Another example which shows the operation of the IM algorithm is shown for the image lighthouse in Figure 6.3. The IM identifies the lighthouse, the building, and the high-contrast crashing waves as the most important areas in the scene. The segmentation did however cause some problems, such as merging the top of the lighthouse with the background, and over-segmenting the strongly textured rocks. An improved segmentation, or the use of the block post-processing technique suggested earlier in this section, would reduce the influence that these segmentation errors have on the final IM. Additional IMs, for a wide range of different scenes, are contained in Appendix A.

A significant characteristic of eye movements and visual attention which was identified in Section 2.3.1.1 is that for typical natural scenes, only a few areas are considered by viewers to be of high importance, while a majority of the scene never attracts our attention or is foveated. An analysis of IMs demonstrates that they also identify only a small area of a scene as being of high importance, while most of the scene is classified as low importance. Table 6.1 gives a breakdown of the distribution of importance generated by the IMs for the 11 test images shown in Appendix A. Although there are some expected differences between the distributions for different scenes, some common patterns are evident. The average
Figure 6.3: IM for the lighthouse image. (a) Original image; (b) segmented image; feature maps for (c) contrast, (d) size, (e) shape, (f) location, (g) foreground/background; and (h) final IM produced by summing (c)–(g). For (c)–(h), lighter regions represent higher importance.
### Table 6.1: Percentage of pixels in each Importance quartile for the 11 test images in Appendix A.

<table>
<thead>
<tr>
<th>Image</th>
<th>Importance Range</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.00–0.25</td>
</tr>
<tr>
<td>Announcer</td>
<td>50.77</td>
</tr>
<tr>
<td>Beach</td>
<td>68.65</td>
</tr>
<tr>
<td>Boats</td>
<td>45.77</td>
</tr>
<tr>
<td>Claire</td>
<td>63.01</td>
</tr>
<tr>
<td>Football</td>
<td>85.15</td>
</tr>
<tr>
<td>Island</td>
<td>57.38</td>
</tr>
<tr>
<td>Lena</td>
<td>29.80</td>
</tr>
<tr>
<td>Lighthouse</td>
<td>67.51</td>
</tr>
<tr>
<td>Miss America</td>
<td>74.07</td>
</tr>
<tr>
<td>Pens</td>
<td>32.12</td>
</tr>
<tr>
<td>Splash</td>
<td>83.52</td>
</tr>
<tr>
<td>Average</td>
<td>59.80</td>
</tr>
</tbody>
</table>

area across all scenes classified as very important (importance range 0.75–1.00) is 7.50%, while on average nearly 60% of the scene are classified as being of low visual importance (importance range 0.00–0.25). This kind of distribution is in agreement with eye movement and attention studies (see Section 2.3.1.1).

### 6.3 Verification of IMs using Eye Movement Data

Visual inspection of the IMs is a useful first step for assessing their accuracy. However, to obtain a quantitative measure of the accuracy of the IMs, a comparison with formal subjective test data needs to be performed. To obtain this data, an eye tracker was used to record the fixations of subjects when viewing the 11 test scenes shown in Appendix A. This section describes the eye tracking experiments, presents their results, and details the correlation between the subject’s eye movements and the IMs.
6.3.1 Description of Eye Movement Experiments

Seven subjects (undergraduate and postgraduate engineering students) participated in the experiment (6 males and 1 female). All reported normal visual acuity. The viewing room was the same one described in Section 5.2.1 for the subjective quality experiments. The viewing conditions were compliant with the Rec. 500 requirements, with the exception that the ambient room lighting was higher than that specified in the standard. However, this should not be significant for the purposes of these experiments.

Eye movements were recorded using an ASL model 210 eye tracker. This device required that subjects wear a pair of spectacle frames, on which is mounted an infra-red sensor assembly for detection of the eye movements. The assembly does not interfere with the subject’s normal field of viewing. A chin rest was used to ensure minimal head movement. The device output the horizontal and vertical co-ordinates of the subject’s point of foveation at 125 Hz, with a reported accuracy of better than ±1.0 deg.

Subjects viewed the screen from a distance of four screen heights, which meant that the screen subtended approximately 17.5 deg horizontally and 14 deg vertically. Following a calibration process, the subjects were shown each of the 11 test scenes for ten seconds in a sequential fashion. A full-field grey screen of five second duration separated consecutive scenes. Subjects were instructed to view the scenes as if they were viewing them for entertainment purposes. Calibration was again performed at the end of the viewing session, to check for any deviations which may have occurred during the session. None of the subjects reported any difficulties in undertaking the experiment, and all data was used for subsequent analysis.
6.3.2 Results of Eye Movement Experiments

The raw eye movement data was first processed to remove any spurious recordings due to the blinking of subjects. The location and duration of fixation points was then determined. A fixation was deemed to have occurred if the point of foveation did not move from within a small local area for a period of at least 100 msec. The fixation locations for all subjects were combined for each image and are shown in Appendix C. These fixations can be compared to the original images and IMs in Appendix A. The typical eye movement patterns of subjects when viewing natural scenes, which were discussed in Section 2.3.1.1, are evident in these fixation plots. There was a strong correlation between the fixation points of different subjects. Although the correlation was strongest for simple scenes (e.g., head-and-shoulders), a high correlation was still observed for complex scenes. Viewers tended to foveate on a few ROIs in the scenes (typically 1–3), rather than scan all areas in the scene evenly. Although the overall destination of fixations was similar for all subjects, the temporal order in which they viewed the different ROIs did not follow any obvious pattern, and was not correlated very strongly across subjects.

6.3.3 Comparison of Fixations and IMs

In order to provide a quantitative comparison of the subject’s fixations with the IMs, the IMs were broken into two parts: the regions which had the highest importance and the regions with the lowest importance. The distinction was made based on the total area of the scene in each of these categories. Two cases were used: one which had 10% of the total scene area in the highest importance category, and one which had 20% of the scene in the highest importance category. It should be noted that in general it is not possible to determine the most important 10% or 20% of the scene with exact precision, since the IMs are measured
in regions while area is measured in pixels. The approach taken was to keep adding the regions of highest importance until at least 10% or 20% of the total image area was classified as being of highest importance. Therefore, the total area classified as being of highest importance will slightly exceed 10% and 20%. These increases in area have been recorded and are stated where appropriate (see the “Actual Area” columns in Table 6.2).

Following this IM categorisation, the location of each fixation could be compared with the IMs to determine whether the fixation occurred in a region classified as highest importance (termed a “hit”), or whether it had occurred in a region of lower importance (termed a “miss”). An example of this process for the scene island is shown in Figure 6.4. It is interesting to note in Figure 6.4(d) the large number of fixations which just miss an important region. This is particularly evident around the trees and near the clouds. This occurs for three main reasons. Firstly, the accuracy of the eye tracker is only ±1.0 deg, so recording may differ slightly from actual fixation points. Secondly, even when fixating on an object, drift eye movements can cause slight changes in the point of fixation (see Section 2.3.1). Thirdly, the segmentation process used in the IM algorithm suffers from some inaccuracy, particularly for small objects. A good way to compensate for these effects is to apply the block post-processing operation on the IMs (described in Section 6.2) prior to a comparison with the fixation data. The operation of this block post-processing for various block sizes can be seen in Figure 6.5. As the block size is increased, the area occupied by the regions of highest importance in the scene also increases. The effect can be thought of as being a simplified version of a morphological dilation operation.

The percentage of hits and misses have been calculated for all scenes at both the 10% and 20% highest importance levels. This has been done for the original IMs and also for the block post-processed IMs, using block sizes of $8 \times 8$, $16 \times 16$, and $32 \times 32$ pixels. The results are shown in Figure 6.6 for the 10% importance
6.3 Verification of IMs using Eye Movement Data

Figure 6.4: Comparison of fixations with IMs for the *island* image. (a) Original image, (b) plot of all fixations, (c) hits for the 10% most important regions (represented by black dots), and (d) misses at the 10% level (represented by white dots).

level, and Figure 6.7 for the 20% importance level. These results indicate several interesting features of the IMs.

- For all scenes, the correlation of the IMs with the fixations was significantly better than would be expected at the chance level (the chance level would be just above 10% and 20% for the two cases).

- Correlation was highest for the simple head-and-shoulders scenes (*announcer*, *claire*, and *Miss America*). Even at the 10% level, correlation with these scenes is high. Increasing to the 20% level generally did not result in a dramatic increase in correlation for these types of scenes, because most of
Figure 6.5: The effect of block post-processing on IMs for the island image. (a) Original IM, (b) IM post-processed with $8 \times 8$ blocks, (c) IM post-processed with $16 \times 16$ blocks, and (d) IM post-processed with $32 \times 32$ blocks.

the fixations were already accounted for at the 10% level.

- For the more complex scenes, a significant improvement in correlation was noticed when moving from the 10% to the 20% levels. This is expected, since the fixations generally occur over a wider area in these more complex scenes, due to the presence of more than one ROI in these scenes.

- The use of block processing often (but not always) produced an increase in the proportion of hits.

To quantify the influence that block post-processing has on the proportion of hits, the results of Figures 6.6 and 6.7 have been averaged across all scenes and are
Figure 6.6: Proportion of hits for each of the test images using the 10% area of highest importance. (a) Original IM, (b) IM post-processed with 8 × 8 blocks, (c) IM post-processed with 16 × 16 blocks, and (d) IM post-processed with 32 × 32 blocks. The order of the images is: 1=announcer, 2=beach, 3=lighthouse, 4=splash, 5=boats, 6=claire, 7=football, 8=island, 9=lena, 10=Miss America, and 11=pens.
Figure 6.7: Proportion of hits for each of the test images using the 20% area of highest importance. (a) Original IM, (b) IM post-processed with $8 \times 8$ blocks, (c) IM post-processed with $16 \times 16$ blocks, and (d) IM post-processed with $32 \times 32$ blocks. The order of the images is: 1=announcer, 2=beach, 3=lighthouse, 4=splash, 5=boats, 6=claire, 7=football, 8=island, 9=lena, 10=Miss America, and 11=pens.
presented in Table 6.2. It can be seen that the block post-processing of the IMs results in an increase in the proportion of hits, when averaged across all scenes. The highest correlation occurs when a $32 \times 32$ block size is used. This table also represents the proportion of hits in terms of the duration of the fixations. In all cases, the proportions when measured using duration of the hits were slightly larger than those measured using the number of hits. This agrees with evidence that subjects will fixate for longer periods of time on objects which attract their interest [112, 262]. Also shown in this table are the actual areas occupied by the highest importance regions, averaged across all scenes. As mentioned earlier in this section, these values will exceed the 10% and 20% values which were targeted. Actual values range from 12.8–15.3% at the 10% level, and 23.3–26.8% at the 20% level. The slight variation in these values across IMs should also be considered when comparing results from the different IMs.
The Importance Map technique for still images has been extended to operate on video sequences. This required the calculation of the motion of objects in the scene, since it is known that motion has a strong influence on our visual attention (see Section 2.3.4.1). A block diagram of the IM technique for video is shown in Figure 6.8. The model requires both the current and previous frames as input. The calculation of the spatial importance occurs in the same way as described in Section 6.2 for still images. The motion importance requires the calculation of the motion of objects in the scene. This is accomplished via a block matching process between the current and previous frames. A simple block matching procedure, such as that suggested for MPEG encoding [170], produces a noisy motion field which does not correspond very well with the actual motion of objects in the scene. To produce a motion field which corresponds more closely
with actual object motion, a technique of successive refinement similar to that suggested by Westen et al. [278] is adopted. Motion vectors are first calculated at a coarse resolution (64 × 64 or 32 × 32 pixels, depending on the size of the image). This is done using a minimum mean absolute difference (MAD) block matching algorithm, weighted slightly to favour zero motion (to reduce noise effects). The resolution of the motion vectors is then increased successively by a factor of two, until a motion vector is obtained at the 8 × 8 pixel level. During this refinement process, the choice of motion vector for a block is constrained to be either its previous value, or the value of one of its four neighbouring blocks (whichever provides the best match in terms of MAD). This successive refinement procedure produces a good estimate of actual object motion, and is robust against noise in the image. It can however fail to detect the motion of very small objects in a scene with complex motion. In such cases, it is best to begin the process with a smaller block size (e.g. 32 × 32 or 16 × 16 pixels).

A motion importance factor $ImpMot$ is then calculated for every block in the frame, using:

$$ImpMot_j = \begin{cases} 
0.0 & \text{if } mot_j < mot_{min}, \\
\frac{mot_j - mot_{min}}{mot_{p1} - mot_{min}} & \text{if } mot_{min} < mot_j < mot_{p1}, \\
1.0 & \text{if } mot_{p1} < mot_j < mot_{p2}, \\
\frac{mot_{max} - mot_j}{mot_{max} - mot_{p2}} & \text{if } mot_{p2} < mot_j < mot_{max}, \\
0.0 & \text{if } mot_j > mot_{max},
\end{cases}$$

(6.7)

where $mot_j$ is the magnitude of the motion vector for block $j$, $mot_{min}$ is the minimum important motion parameter, $mot_{p1}$ and $mot_{p2}$ are peak important motion parameters, and $mot_{max}$ is the threshold for maximum important motion. This is depicted graphically in Figure 6.9. High importance is assigned to regions undergoing medium to high motion, while areas of low motion and areas undergoing very high motion (i.e. objects with untrackable motion) are assigned low motion.
importance. The optimal choice of model parameters can vary depending on the nature of the motion of the scene. However, good results have been achieved across a broad range of scenes undergoing various degrees of motion using parameter values of $mot_{\text{min}} = 0.0 \text{ deg/sec}$, $mot_{p1} = 5.0 \text{ deg/sec}$, $mot_{p2} = 10.0 \text{ deg/sec}$, and $mot_{\text{max}} = 15.0 \text{ deg/sec}$.

Once again there is little quantitative data to suggest an optimal weighting of the spatial and motion IMs to produce an overall IM. Therefore, the average of the two is taken. This equal weighting of spatial and motion IMs emphasises the strong influence of motion on visual attention, since the spatial IM is constructed using five factors while the motion IM only uses one. A similar weighting of spatial and motion factors was recommended by Niebur and Koch [179] in their attention model. They weight motion as having five times the influence of any other factor, based on their own empirical findings.

Figure 6.10 shows the results of the video IM calculation for a frame of the football sequence. The original scene is quite complex, with several objects undergoing large amounts of motion. The spatial IM (Figure 6.10(b)) has identified the players as the parts of the scene with highest visual importance. The motion IM (Figure 6.10(c)) classifies the importance of objects based on their velocity.
Figure 6.10: Importance Map for a frame of the football sequence. (a) Original image, (b) spatial IM after $8 \times 8$ block post-processing, (c) motion IM, and (d) overall IM at $16 \times 16$ resolution. Lighter regions represent higher importance.

Note the extremely high motion in the upper body of the player entering at the right of the frame. The motion IM has classified this region as being untrackable by the eye, and it is therefore assigned low importance. The final IM shown in Figure 6.10(d) was produced by averaging the spatial and motion IMs. It gives a good indication of the visually important regions in the frame. The IM is shown at macroblock resolution ($16 \times 16$ pixels), which makes it amenable for use in an MPEG coding algorithm (see Chapter 8).

The results of the IM calculation for a frame of the table tennis sequence are shown in Figure 6.11. Once again, the visually important regions have been well
Figure 6.11: Importance Map for a frame of the table tennis sequence. (a) Original image, (b) spatial IM after 8×8 block post-processing, (c) motion IM, and (d) overall IM at 16 × 16 resolution. Lighter regions represent higher importance.

identified by the IM. In this particular scene, the spatial IM did not identify the ball due to a coarse segmentation. However, by combining the spatial and motion IMs, the overall IM was successful in identifying the ball as an important part of the scene. This is a good example of the benefits which can be obtained by using multiple factors in the calculation of IMs.

In the next chapter, the utility of the IM algorithm is demonstrated by combining it with the early vision model of Chapter 5 to produce an attention-weighted image quality metric.
A new early vision model for image fidelity and quality assessment was presented in Chapter 5. An analysis of this model’s performance showed that it gives accurate predictions of the fidelity of compressed natural images. To convert the fidelity metric to a quality metric, a simple Minkowski summation was performed across all visible distortions. Correlation with subjective MOS indicated that, although this quality metric provided significant improvement over simple metrics like MSE, there was still scope for improvement, particularly in scenes with suprathreshold distortion or with spatially-varying quality. Increasing the complexity of the early vision model only resulted in minor changes in prediction accuracy. Thus it appears that an early vision model by itself is not powerful enough to provide accurate predictions of picture quality over a broad range of scenes, bit-rates, and compression techniques. This is because these models do not take into account any higher level attentional or cognitive effects, which are known to have a strong influence on subjective picture quality.
One effect that needs to be considered is that distortions have a different impact on picture quality depending on their location with respect to the scene content. A review of eye movements and visual attention in Chapter 2 revealed a strong correlation between the eye movements of different subjects when viewing natural scenes. A number of bottom-up and top-down factors which influence visual attention and eye movements were identified. These findings provided the motivation for the Importance Map model in Chapter 6. The IM technique is designed to detect ROI in a natural scene in an unsupervised manner. Correlation with eye movements verifies the utility of this model.

In this chapter, the early vision model of Chapter 5 is combined with the Importance Map model to produce a quality metric which takes into account the distortion's location with respect to scene content. This is performed by weighting the PDMs (produced by the early vision model) with the IMs, prior to a summation of the distortions. The model is described in Section 7.1. The performance of the new metric is quantified in Section 7.2, by correlating the model's predictions with subjective test data. This section also examines the influence of different model parameters and configurations.

### 7.1 Model Description

A block diagram of the new quality assessment technique is shown in Figure 7.1. The overall operation is similar to the models described in [188, 192]. The multi-channel early vision model is described in detail in Section 5.1. In brief, it consists of the following processes: a conversion from luminance to contrast and a multi-channel decomposition using Peli’s LBC algorithm [195]; application of a CSF using data obtained from naturalistic stimuli; a contrast masking function which raises the CSF threshold more quickly in textured (uncertain) areas than in flat or edge regions; visibility thresholding; and Minkowski summation. This process
produces a PDM which indicates the parts of the coded picture that contain visible distortions, assuming that each area is in foveal view. This map gives a good indication of the fidelity of the picture.

As discussed in Section 4.2.2, our perception of overall picture quality is influenced strongly by the fidelity in the ROIs in a scene. The IM algorithm is able to detect the location of ROIs in a scene. A complete description of the IM algorithm is contained in Section 6.2. In brief, the original image is first segmented into homogeneous regions. The salience of each region is then determined with respect to each of five factors: contrast, size, shape, location, and foreground/background. The factors are weighted equally and are summed to produce an overall IM for the image. Each region is assigned an importance value in the range from 0.0–1.0, with 1.0 representing highest importance.

The IMs are used to weight the output of the PDMs. Although it is known that some areas in a scene are more important than others, little data exists regarding the quantitative weighting of errors with respect to their location and magnitude. For example, it is not known whether a distortion which is 3.0 times the threshold of visibility in a low importance region is as objectionable as a distortion that is 1.5 times the visibility threshold in an area of high importance. These kinds of relationships are likely to be very complex and non-linear, and may be dependent on the specific scene and application.
on a number of factors (e.g., eccentricity of viewing, magnitude and structure of distortion); hence it is difficult to design experiments to obtain data which has general applicability. In the absence of suitable quantitative data, a simple approach for weighting the PDM is adopted:

\[ IPDM(x, y) = PDM(x, y) \cdot IM(x, y)\gamma, \]  

(7.1)

where \( IPDM(x, y) \) represents the IM-weighted PDM, and \( \gamma \) is an importance scaling power. Different values of \( \gamma \) have been tested on the model and its correlation with subjective data determined (see Section 7.2.2 for an analysis). Best correlation with subjective MOS was obtained using \( \gamma = 0.5 \) when the raw IM was used. However, the optimal value of \( \gamma \) increased slightly when the IM was post-processed (see Section 7.2.2).

To combine the IPDM into a single number, Minkowski summation is performed over all pixels \( N \):

\[ IPQR = \left( \frac{1}{N} \sum_{N} IPDM(x, y)^\beta \right)^{1/\beta}, \]  

(7.2)

where \( IPQR \) represents the IM-weighted Perceptual Quality Rating. The value of \( \beta \) is used to control the output. A value of \( \beta = 1.0 \) would result in a quality proportional to the average of errors over all pixels, while for high values of \( \beta \), the quality becomes proportional to the maximum IPDM value. As was the case in the summation of the early vision model in Section 5.1.5, a value of \( \beta = 3.0 \) was found to give most accurate results. This is in agreement with the subjective data of Hamberg and de Ridder [103]. However, overall correlation did not change significantly if \( \beta \) was increased to 4.0.

Finally, to produce a result on a scale from 1–5 and to enable direct comparison
with MOS data, the summed error can be scaled as

\[ IPQR_{1-5} = \frac{5}{1 + p \cdot IPQR}, \]  

(7.3)

where \( IPQR_{1-5} \) represents the IPQR scaled to the range \([1.0-5.0]\), and \( p \) is a scaling constant. Following subjective testing, a value of \( p = 0.8 \) was found to provide good correlation with subjective opinion.

### 7.2 Performance of the Quality Metric

A similar methodology to that used in Section 5.3 (for assessing the early vision model) was used to determine the accuracy of the new quality metric. Firstly, the IPDMs produced by the model were inspected and compared to the PDMs. The aim of this inspection was to check that the IM-weighting of the PDMs was producing the desired effect of reducing the magnitude of distortions in less important areas in the scenes, and that this reduction was in line with subjective perception of quality. As a first step, this enabled an informal, qualitative evaluation of the quality metric. Secondly, the correlations of the model’s IPQR predictions with the MOS data obtained from the subjective tests in Section 5.2 were calculated and compared to the correlations achieved by the PQR metric.

The quality metric has been tested using a wide range of different scenes, bitrates, and coders. A typical example of its operation can be seen in Figure 7.2 for the image *island*, JPEG coded at 0.95 bit/pixel. This coded scene has a very high subjective quality, since the detectable distortions are small and are not located in the ROIs. The IM shown in Figure 7.2(c) gives an indication of the location of ROIs in the scene. The problems associated with MSE-based metrics are evident in Figure 7.2(d). The squared error overemphasises the visual distortion in textured areas, since it doesn’t account for the strong masking which occurs in these areas. The PDM in Figure 7.2(e) identifies the location of visible
Figure 7.2: Outputs of the quality metric for the image *island*, JPEG coded at 0.95 bit/pixel. (a) Original image, (b) coded image, (c) Importance Map, with lighter regions representing areas of higher importance, (d) squared error, (e) PDM, and (f) IPDM.
Figure 7.3: Outputs of the quality metric for the image *island*, JPEG coded at 0.50 bit/pixel. (a) Coded image, (b) squared error, (c) PDM, and (d) IPDM.

errors, under the assumption that the whole scene is being viewed foveally. It represents the fidelity of the image. The IPDM in Figure 7.2(f) has weighted the PDM according to the importance assigned by the IM. This provides a means of assessing the objectionability of artifacts, depending on where they occur in an image. Figure 7.3 shows results for the same scene JPEG coded at 0.5 bit/pixel. The IM generated for this image is the same as that shown in Figure 7.2(c). The influence of distortions occurring in the water or in the sky are reduced, but distortions on the island, trees, or clouds maintain their significance.

A good way to demonstrate the utility of this attention-based quality metric is to use a scene coded with spatially varying quality. Figure 7.4 provides a simple
Figure 7.4: Outputs of the quality metric for the image *island*, JPEG coded at two levels for an average bit-rate of 0.58 bit/pixel. (a) Coded image, (b) squared error, (c) PDM, and (d) IPDM.

example of this. The coded scene is a composite of Figures 7.2(b) and 7.3(a). Rectangular areas corresponding to the ROIs in the scene have been cut from the high quality Figure 7.2(b) and pasted onto Figure 7.3(a), in the manner described in Section 5.2.2. The result is an image with high fidelity in the ROIs, and lower fidelity in the periphery. This type of image typically has a higher subjective quality than an image of the same bit-rate which has been coded at a single quality level. However, if the squared error, PDM, and IPDM are each combined into a single number representing image quality, only the IPDM is capable of detecting the increased subjective quality. This can be seen in Figure 7.5, which shows the IPQR, PQR, and PSNR for *island* at a range of bit-rates and using both wavelet and JPEG encoding. The variable quality images are assigned qualities
the same as or lower than the single quality images by both the PSNR and PQR metrics. However, the IPQR metric assigns a higher quality to the scenes coded with variable quality, which is in accordance with human perception. This type of result is seen again later in this section, when the quality metric is compared with subjective MOS data.

As a further example, Figures 7.6 and 7.7 show the result of using the new metric with the image *announcer*. In Figure 7.6, the coded image was generated using a combination of high and low bit-rate wavelet coded images, in a similar fashion to
Figure 7.6: Outputs of the quality metric for the image *announcer*, wavelet coded at two levels for an average bit-rate of 0.37 bit/pixel. (a) Original image, (b) coded image, (c) IM, (d) squared error, (e) PDM, and (f) IPDM.
7.2 Performance of the Quality Metric

Figure 7.7: Predicted qualities for the image *announcer*. (a) PSNR, (b) PQR, and (c) IPQR. '*' = uniform JPEG; '+' = uniform wavelet; '□' = variable JPEG; '○' = variable wavelet.

the coded image in Figure 7.4(a). Once again, the IPQR is the only method which predicts an increase in quality for this image in comparison to a homogeneously coded image of the same bit-rate, as shown in Figure 7.7.

The correlation of the PQR with the subjective test data of Section 5.2 has been determined and is plotted in Figure 7.8. The correlation of the PQR with subjective MOS (from Figure 5.9) has been included for comparison. The IPQR technique achieves a much better correlation than PQR over the broad range of subjective qualities tested. This is most apparent for the scenes coded with
Figure 7.8: Correlation of objective quality metrics with subjective MOS. (a) IPQR, and (b) PQR. ‘○’ = Uniform JPEG; ‘×’ = Uniform wavelet; ‘□’ = Variable JPEG; ‘◦’ = Variable wavelet.
variable quality. The PQR typically underestimates the quality in these types of scenes, since it fails to take into account the location of errors with respect to the scene content. However, by using the IMs, the IPQR is able to provide a good prediction for these challenging scenes. Another interesting result is the strong correlation across scenes and coders for very high quality scenes (i.e. MOS≈5.0). By taking the location of distortions into account, the IPQR is also able to provide a better estimate of the point where distortions begin to influence subjective quality.

The overall correlations with subjective MOS achieved by these techniques are shown in Table 7.1. Across all scenes (both uniform and variable coded), IPQR achieves a correlation of \(|r| = 0.93\) with MOS, an improvement from the \(|r| = 0.87\) of PQR. As expected, the improvement in correlation of the IPQR is strongest for the variable coded scenes. However, there is still a noticeable improvement in correlation for uniform coded scenes (from \(|r| = 0.94\) to \(|r| = 0.97\)). This indicates that the use of IMs in quality metrics can be beneficial for uniform as well as variable-quality coded scenes.

To test the statistical significance of the improvement in correlation achieved by IPQR (\(|r| = 0.93\)) over PQR (\(|r| = 0.87\)), the hypothesis \(H_0 : \rho_1 = \rho_2\) is tested using the Fisher Z independent correlation test from Equation 5.14. This produces a value of \(z = 1.47\) for the difference in correlation of IPQR and PQR, which is significant at the level \(\alpha = 0.10\) but not at the level \(\alpha = 0.05\) \((z_{\alpha=0.10} = 1.28; z_{\alpha=0.05} = 1.64)\). Thus the null hypothesis \(H_0 : \rho_1 = \rho_2\) can be
Table 7.2: Correlation of IPQR with subjective MOS, for different post-processing block sizes.

rejected at the level $\alpha = 0.10$, but not at the $\alpha = 0.05$ level. The improvement in correlation is not statistically significant at the $\alpha = 0.05$ level.

### 7.2.1 Influence of IM Block Post-processing on the Quality Metric

A potential problem can occur when raw IMs are used in a quality metric. The sharp discontinuities in importance at the edges of objects may result in an underestimation of distortions very close to important objects. This is significant, since many coding techniques introduce large distortions near edges. This problem is magnified by segmentation inaccuracies and eye movements, as discussed in Section 6.3.3.

A way to alleviate this problem is to first perform block post-processing on the IMs, as described in Section 6.2. In Section 6.3.3 it was found that post-processing the IMs produced a better correlation with subject’s fixations. The block post-processing operation has also been performed on the IMs used in the quality metric, prior to weighting the PDM. The resulting correlations with MOS are shown in Table 7.2. These results were all obtained using $\gamma = 0.5$. For the $8 \times 8$ and $16 \times 16$ block processed IMs, the correlation across all images, and the correlation with variable-coded scenes, increased in comparison to when raw IMs
were used. The gains obtained by using block post-processing became smaller for 32×32 blocks, but the results were still better than when raw IMs were used. The statistical significance of the improvement in correlation achieved by the 8×8 block processed IPQR in comparison to the PQR was measured, again using the Fisher Z correlation test. The improvement in correlation across all scenes is significant at the α = 0.01 level (z = 2.50; zα=0.01 = 2.33), so the null hypothesis can be rejected at this level. Therefore, the improvement in correlation that the 8×8 block processed IPQR achieves over PQR is statistically significant at the α = 0.01 level.

### 7.2.2 Influence of the Importance Scaling Power on the Quality Metric

To determine the optimal importance scaling power γ, the value of γ was varied over a range and the effect this had on the overall correlation between the IPQR and subjective MOS was measured. The results are shown in Figure 7.9. When raw IMs are used, the correlation rises quite sharply to a definite peak at γ = 0.5. The correlation then tapers off more slowly at higher values of γ. The optimal value of γ increases slightly when block processed IMs are used, to a peak at around γ = 0.7–0.8. This slight increase in the optimal γ is understandable, since the block processing of the IMs reduces the sharp discontinuity in importance near the border of regions. The influence that the areas of low importance have on quality can then be further reduced by increasing the value of γ.
7.2 Performance of the Quality Metric

Figure 7.9: Relationship between the correlation of IPQR and MOS and the importance scaling power $\gamma$. (a) Using raw IMs, (b) using $8 \times 8$ block post-processed IMs, (c) using $16 \times 16$ block post-processed IMs, and (d) using $32 \times 32$ block post-processed IMs.
Chapter 8

Application of Vision Models in Compression Algorithms

In Chapters 5-7 a model of the HVS, based on both early vision and attentional processes, was used in a quality metric for compressed natural images. A high correlation with subjective test data was demonstrated, indicating the accurate and robust nature of this model. Although the primary purpose for the vision model in this thesis was for use in a quality metric, it is important to realise that the utility of the model can be extended to many other application areas. One area which is currently being actively investigated is the incorporation of full or partial human vision models into compression algorithms. This is commonly referred to as perceptual coding.

Although the same basic principles are used in all perceptual coders, the way in which human visual processes are implemented can vary significantly, depending both on the compression technique being used and the vision model which is being incorporated. Some vision models produce outputs which can be used directly by a coding algorithm, while others need to be modified or broken up into components before they can be implemented in a coding algorithm. Vision
models which provide spatially and/or temporally localised information regarding distortions or scene content are particularly amenable to use in a perceptual coding algorithm, since many coders allow for local adaptation. This means that the PDMs and IMs, produced by the early vision and attention models respectively, are suitable for direct use by many different coding techniques.

This chapter describes ways to use early vision and attention models in perceptual coders. Section 8.1 provides a general discussion of the principles and techniques used by perceptual coders, and reviews ways in which properties of the HVS have been applied in compression standards. To provide an example of how the vision models presented in this thesis can be used in a coding algorithm, an MPEG encoder is used. The spatially-adaptive quantisation of the MPEG encoder is controlled by both a simplified early vision model and the IM algorithm. Details of the model are contained in Section 8.2, while verification of the model’s performance through subjective quality testing is provided in Section 8.3. These tests demonstrate that a substantial improvement in subjective quality can be achieved by using vision models in a coding algorithm, even if the model is quite simple.

8.1 Fundamental Principles and Techniques of Perceptual Coders

During the 1970s and early 1980s, a number of new compression techniques were developed. This trend has slowed recently, and fundamental compression technologies are now mature. Furthermore, the worldwide adoption of DCT-based compression standards means that for any new technique to gain broad acceptance, it must be proven to outperform the current standards by a considerable amount. These factors have resulted in a shift in the focus of the compression
research community, to the investigation of ways to tune existing compression techniques rather than the development of fundamentally new techniques.

Perceptual coding is an area which can provide significant improvement in subjective picture quality over conventional coders, and it has recently become an area of active research. Although traditional compression techniques have coarsely modeled some HVS properties (see Section 8.1.1), they have generally been optimised in terms of objective metrics such as MSE. Since it is well known that these simple objective metrics do not correlate well with human perception, better subjective performance can be obtained if the compression techniques are optimised for human visual properties.

The general aim of a perceptual coder can be seen in Figure 8.1. This figure shows an idealised distribution of distortions (in terms of their perceptual visibility) for two types of scenes: those which have been perceptually coded, and those that have not been perceptually optimised. For the perceptually lossless case (Figure 8.1(a)), the perceptually coded scene has distortions which are just
below the threshold of visibility. This allows maximal compression while still maintaining the transparency of the coding process. In Figure 8.1(b), although the distortions are suprathreshold, their tight distribution in the perceptually coded scene ensures that none of the distortions are particularly severe. This ensures maximal compression for a given picture quality, since it is known that our overall perception of picture quality is dependent on the worst areas of distortion in the scene [48, 142].

8.1.1 Perceptual Features Inherent in Compression Standards

The international compression standards JPEG [200, 266], MPEG [139, 170], and H.26x [147] were briefly reviewed in Section 3.1.3. Although the processes are sometimes coarsely modeled, these standards implicitly allow for many properties of the HVS and of natural images in their structure. These are outlined below.

- **Quantisation Matrix (QM).** The default QMs which have been specified for JPEG, MPEG, and H.26x are intended to model the spatial frequency sensitivity of the HVS.

- **DCT block size.** In terms of information theory, better compacting of energy can generally be obtained using a DCT with larger block size. However, the non-stationary structure of natural scenes, along with computational complexity issues, make block sizes of $8 \times 8$ or $16 \times 16$ preferable.

- **Frame rate.** Optimal frame rates have been chosen, so that they are the lowest possible without introducing any perceptible flicker or loss of continuity.

- **Chrominance sub-sampling.** Due to the reduced sensitivity to chrominance of the HVS, the chroma are generally sub-sampled as a pre-processing step.
• **Spatially-adaptive quantisation.** In order to allow for spatial masking effects, MPEG and H.26x allow spatially-varying quantisation by scaling the QM. For MPEG, the recommended way to control this process is specified in the Test Model 5 (TM5) rate control algorithm [239], which crudely approximates spatial masking in the HVS.

• **Scene changes.** Scene changes generally require a large number of bits to code, and induce significant distortion for a few frames following the scene change. However, temporal masking effects reduce visual sensitivity near a scene cut, so that these distortions are typically not perceptible.

Although the compression standards specify default methods for implementing these features, they have purposely been designed in an open manner. Only the actual bit-streams have been specified, which enables encoder designers to implement their own algorithms. This is particularly evident in the MPEG standards. The two areas which are most commonly targeted by perceptual coding algorithms are the design of the QM, and the control of the spatially-adaptive quantisation (see Section 8.1.2.2).

### 8.1.2 Summary of Common Perceptual Coding Techniques

This section summarises the common techniques used to incorporate HVS properties into a coder. A full review and description of all the perceptual coders which have been proposed is beyond the scope of this thesis. Instead, a description of general strategies which have been used is provided, along with references to some of the best known and most relevant exponents of these techniques.

One way of classifying perceptual coding strategies is by the position in which they are implemented within the coding process. This leads to three distinct areas where human perception can be incorporated into a coder:
• *perceptual pre-processing*, which involves modifying the scenes prior to input to later stages of the coder,

• *perceptual quantisation*, characterised by controlling the quantisation process by a (possibly simplified) human vision model,

• *perceptual post-processing*, which involves modifying the coded scenes after they have been compressed, to remove or reduce any visible distortions which were introduced.

### 8.1.2.1 Perceptual Pre-processing

The aim of perceptual pre-processing techniques is to remove any content from the scene which is either imperceptible, or not necessary for the purposes of the application. High frequency noise is present in many scenes, but can be either imperceptible, or may actually reduce the subjective quality of the scene [83]. High frequency noise is also notoriously difficult to encode. For these reasons, several perceptual models of varying complexity have been proposed to remove high frequency content from the scene without compromising its quality (e.g. [6, 7, 260]). Care has been taken in these algorithms not to blur the edges of objects in the scene.

At lower bit-rates, and particularly in situations where prior knowledge of the scene content or viewer’s intentions is available, it is useful to pre-process the scene to remove content in areas which are not ROIs. Performing this at the pre-processing stage is often preferable to using spatially-variable quantisation at a later stage. This is because pre-processing generally blurs or removes content, which typically produces a better quality picture than a coarsely quantised image which may introduce unnatural artifacts such as blockiness. Several models have been proposed which significantly reduce the bandwidth of the image by reducing resolution in areas which are not ROIs [17, 64, 96, 136, 230]. In these models,
the ROIs have either been determined using an eye tracker, or manually by user selection.

### 8.1.2.2 Perceptual Quantisation

The quantisation process has been the area where the majority of perceptual coders have introduced HVS properties into the coding algorithm. The actual way in which this has been implemented depends strongly on the type of coding algorithm which is being used. DCT-based coders have used perceptual coding in two main areas: the design of the QM (to model the CSF), and the control of spatially-adaptive quantisation (to model spatial masking). Watson et al. [201, 270] have proposed an iterative way to calculate image-dependent QMs. In this approach, a scene is first coded using a default QM. Following luminance and contrast masking stages, coding errors are summed and determined in each spatial frequency band. This information is then used in an iterative manner to weight the QM co-efficients, and continues until a perceptually-flat distortion is produced. This technique has recently been extended to medical images by Eckert and Jones [73].

Many techniques have been proposed which use a human vision model to control the spatially-adaptive quantisation process in an MPEG or H.26x encoder. This is important, since the TM5 and RM8 models suggested by MPEG and H.261 respectively are very simple and do not model masking effects very closely. One approach has been to categorise the scenes into flat, edge, and textured regions, and then to allow higher quantisation in the textured areas (e.g. [98, 145, 235]). These methods have demonstrated improved subjective quality, in comparison to video which was MPEG coded using TM5. Another more computationally expensive approach is to iteratively adjust the spatially-varying quantisation after running the output through an early vision model [263, 264]. Areas of high distortion will have the amount of quantisation reduced, while those areas which are below the
threshold of visibility will undergo increased quantisation. An alternative use of this early vision model approach has been proposed by Westen et al. [278]. Their model tracks moving objects, so the acuity of objects which are being tracked by the eye is not reduced. Unfortunately no subjective comparison with TM5-coded sequences has been reported for either of these models. Other spatially-adaptive quantisers have been proposed which increase quantisation in areas which are not ROIs [157]. This has been particularly useful in video-conferencing applications, where the ROIs can be automatically detected [36, 221, 287]. However, care must be taken not to introduce strong peripheral distortions, which may distract a viewer’s attention [66].

Properties of the HVS have also been used to control the quantisation processes of wavelet [115, 183], subband [37, 218], and fractal [144] coders. It could also be argued that region-based [137] and model-based [4] coders inherently incorporate human visual properties, since they segment the scene into natural objects in a similar way to the HVS.

8.1.2.3 Perceptual Post-processing

Although it is desirable not to introduce visible distortions in the encoded scene, this is not always possible due to bandwidth and encoder limitations. One way that a decoder can improve the subjective quality of such pictures is to employ perceptual post-processing. This aims to reduce or remove any visually noticeable coding errors. In general, some knowledge of the type of coder which was used to code the scene is required. For example, block-based coders are known to introduce distortions around block edges when operated at low bit-rates. This knowledge, combined with simple masking or classification models which take care not to remove edges in the scene, have recently been used to provide improved subjective picture quality in block-based coders [165, 260, 288].
8.2 A New MPEG Quantiser Based on a Perceptual Model

As discussed earlier in this section, the MPEG standards have been designed in open manner, with only the bit-stream being specified. This flexibility enables the designers of coding algorithms to include their own algorithms in many areas of the coding process (see Section 8.1.2). The most open area is the control of the quantisation, and it is this area which is targeted by most perceptual MPEG coders. In this section the spatially-adaptive quantisation process of MPEG is briefly reviewed, and a new adaptive quantiser is described. This algorithm uses both a simplified early vision model and the IM technique to control quantisation. Some results of frames coded using this technique are shown, for comparison with TM5 coded frames. Although the new quantiser has been implemented using an MPEG-2 encoder [70], the algorithm could readily be applied to an MPEG-1 or H.26x encoder. Since the basic operation of the MPEG-1 and MPEG-2 standards is similar, they are referred to generically as MPEG unless otherwise specified.

8.2.1 MPEG Adaptive Quantisation

MPEG is a block-based coding method based on the DCT. A sequence is made up of groups of pictures, which contain pictures of three different types: I- (intra-coded), P- (predictive-coded) and B- (bi-directionally predictive-coded). Each picture is broken into $16 \times 16$ pixel regions (macroblocks), which are further broken into $8 \times 8$ pixel regions before a DCT transform is applied. This transform significantly reduces the spatial redundancy in the frame and allows efficient representation following quantisation via a user-defined QM. The QM is designed both on the statistics of natural images, and on the spatial frequency sensitivity of the HVS. The significant temporal redundancy is reduced by using predictive P- and B- pictures, as well as independent I- pictures. Thus, through use of
motion compensation in P- and B- pictures, only differences in adjacent pictures need to be coded. For a complete description of the MPEG standard, refer to Mitchell et al. [170].

The quantisation strategy used in MPEG revolves around the QM. Although a new QM may be specified only once per picture, it would be beneficial to use spatially-variable quantisation within a picture. This is because some areas of the picture can tolerate more severe distortion than others, due primarily to masking and attentional factors. MPEG has allowed for this via the MQUANT parameter, which allows a scaling of the QM (by a factor of 1 to 31) for each macroblock. MPEG-2 provides further flexibility by allowing a non-linear scaling of the QM via MQUANT.

The MPEG-2 committee developed a strategy for the usage of MQUANT in their TM5 controller [239]. This however was designed only as a basic strategy, and a more complex technique can easily be adopted. TM5 involves three basic processes:

- target bit allocation for a frame,
- rate control via buffer monitoring,
- adaptive quantisation based on local activity.

The adaptive quantisation is the area which is targeted by the new model described in Section 8.2.2. The adaptive quantisation strategy used in TM5 crudely models our reduced sensitivity to complex spatial areas (i.e. spatial masking) by varying MQUANT in proportion to the amount of local spatial activity in the macroblock. The spatial activity measure used in TM5 is the minimum variance among the four $8 \times 8$ luminance blocks in each macroblock. However, this simple activity measure fails to take into account such HVS properties as our different sensitivities to edges and textures, and attentional factors. For these reasons, a
Figure 8.2: Block diagram for MPEG adaptive quantisation controller.

more complex model is required to provide compressed pictures with improved subjective quality.

8.2.2 Description of New Quantiser

The method of adaptive quantisation detailed here is similar to that described in previous publications [189, 191]. The basic operation can be seen in Figure 8.2. The algorithm has been developed based on the spatial masking and visual attention properties of the HVS discussed in Sections 2.2.3 and 2.3 respectively. It is intended to replace the adaptive quantisation strategy used by TM5, and has been designed to have low computational expense. The original frame is input, along with the previous frame for motion calculation. In the current implementation, only luminance frames are used, but chrominance factors may readily be incorporated. The upper branch of Figure 8.2 shows the spatial masking strategy. The image is broken up into $8 \times 8$ blocks, and each block is classified as either flat,
edge, or texture. This is performed using the same technique as that used in the early vision model, and is described in Section 5.1.4. Activity is then measured in each block by computing the variance (as done in TM5). This activity value is adjusted based on the block classification as follows:

\[ act'_j = \begin{cases} 
\min(\text{act}_j, \text{act}_{th}) & \text{if region is flat}, \\
\text{act}_{th} \left( \frac{\text{act}_j}{\text{act}_{th}} \right)^\varepsilon & \text{if region is edge or texture},
\end{cases} \tag{8.1} \]

where \( act'_j \) is the adjusted activity for block \( j \), \( \text{act}_j \) is the variance of block \( j \), \( \text{act}_{th} \) is the variance visibility threshold, \( \varepsilon = 0.7 \) for edge areas, and \( \varepsilon = 1.0 \) for textured areas. A value of \( \text{act}_{th} = 5.0 \) has been found to provide a good estimate of the variance allowable in a flat block before distortion becomes visible. The adjusted activity is then used to control MQUANT as in TM5:

\[ N\text{act}_j = \frac{2\text{act}'_j + \text{act}_{avg}}{\text{act}'_j + 2\text{act}_{avg}}, \tag{8.2} \]

where \( N\text{act}_j \) is the normalised activity for block \( j \), and \( \text{act}_{avg} \) is the average value of \( \text{act}'_j \) for the previous frame. \( N\text{act}_j \) is now constrained to the range \([0.5-2.0]\). This technique ensures minimal quantisation error in flat regions, increases quantisation with activity gradually along edges, and increases quantisation with activity more significantly in textured regions.

The incorporation of IMs into the coder can be seen in the lower branch of Figure 8.2. The IM technique used is the same as that described in Section 6.2 for still images, with the extension to video of Section 6.4. The spatial and motion importance maps are combined by averaging. Since a value for controlling MQUANT is only required for every 16 × 16 pixel macroblock, the highest importance value in the four 8 × 8 blocks constituting the macroblock is used as the macroblock importance value. This factor is used to control the adaptive
quantisation in a similar manner as that used by the spatial masking process:

\[ N_{imp_j} = \frac{imp_j + 2imp_{avg}}{2imp_j + imp_{avg}}, \quad (8.3) \]

where \( N_{imp_j} \) is the normalised importance for block \( j \), and \( imp_{avg} \) is the average value of \( imp_j \) for the previous frame. The final value of MQUANT is calculated using the results of both the spatial masking and importance maps:

\[ MQUANT_j = N_{act} \cdot N_{imp_j} \cdot Q_j, \quad (8.4) \]

where \( Q_j \) is the reference quantisation parameter calculated by the MPEG rate control procedures.

### 8.2.3 Typical Results Achieved Using New Quantiser

The perceptual quantiser has been tested on a wide variety of sequences. The results have been promising, even for relatively complex scenes. The activity masking model provides better subjective quality than TM5's activity measure, since distortions are transferred from edge areas to texture areas, which can tolerate more error due to larger masking effects. This is demonstrated by a reduction in ringing near object edges. The incorporation of IMs in the adaptive quantisation process provides further subjective improvement, since distortions are now reduced in visually important regions. This is particularly evident at low bit-rates.

Although it is difficult to demonstrate video quality by showing still frames, some coding results which compare the new technique to the standard TM5 rate controller have been included. Low bit-rates have been used in order to accentuate the artifacts. Figure 8.3 shows the results using the same bit-rates for a frame of the Miss America sequence. Severe blocking can be seen in the visually im-
8.2 A New MPEG Quantiser Based on a Perceptual Model

![Figure 8.3: Coding results for a frame of the Miss America sequence at 350 kbit/sec. (a) Original frame, (b) frame coded using the new perceptual quantiser, and (c) frame coded using TM5.](image)

important facial regions in the TM5 coded scene. However, the results for the new coder show reduced distortion in the visually important areas. This can be seen more clearly in Figure 8.4, which shows the same scenes zoomed in around the face. For a frame from the table tennis scene (Figure 8.5), the TM5 coder has created ringing artifacts around the hand, arm, racquet, and the edge of the table. Blocking effects can also be seen in the hand. These distortions are reduced using the new technique, at the expense of increased degradation in the (lower importance) background. A clearer indication of the reduction of the artifacts
Figure 8.4: Coding results for a frame of the Miss America sequence at 350 kbit/sec, zoomed in around the face. (a) Original frame, (b) frame coded using the new perceptual quantiser, and (c) frame coded using TM5.
Figure 8.5: Coding results for a frame of the *table tennis* sequence at 1.5 Mbit/sec. (a) Frame coded using the new perceptual quantiser, and (b) frame coded using TM5. The original frame is shown in Figure 6.11(a).
Figure 8.6: Coding results for a frame of the table tennis sequence at 1.5 Mbit/sec, zoomed in around the racquet. (a) Frame coded using the new perceptual quantiser, and (b) frame coded using TM5.

around the hand and racquet can be seen in the close-up view in Figure 8.6.

8.3 Performance of the Perceptually-Based MPEG Quantiser

To quantify the performance of the new perceptual quantiser in comparison to standard TM5, subjective assessment of the compressed video sequences was performed. The stimulus comparison methods of Rec. 500-6 were chosen for this
8.3 Performance of the Perceptually-Based MPEG Quantiser

purpose (see Section 3.2). Twelve subjects (undergraduate and postgraduate engineering students) participated in the tests. All reported normal visual acuity. The viewing room was the same one described in Section 5.2.1 for the subjective quality experiments. The viewing conditions were compliant with the Rec. 500 requirements.

Two different sequences were used in the test: Miss America and football. A single frame from these sequences is shown in Figures A.5 and A.9. Both sequences were approximately five seconds in length. The Miss America sequence is a head-and-shoulders view of a woman who is talking. It is quite simple, with limited motion. The football sequence however is taken from an American football game, and contains large amounts of motion. It is therefore quite difficult to code. To obtain the test stimuli, each sequence was coded at two bit-rates using the perceptual encoder. For Miss America these bit-rates were 250 kbit/sec and 300 kbit/sec, while for football they were 500 kbit/sec and 700 kbit/sec. The TM5 coded sequences were obtained at several bit-rates which were equal to or greater than the bit-rates of the perceptually coded sequences. By doing this, the bit-rate at which the subjective quality of the TM5 coded sequence equaled the subjective quality of the perceptually coded sequence could easily be determined.

The testing strategy used was to ask subjects to compare the quality of TM5 and perceptually coded sequences. The adjectival categorical judgment method of Rec. 500-6 was used as the scale. The original and coded scenes were shown in successive pairs, according to the timing diagram shown in Figure 8.7. The order in which the original and coded scenes were presented was pseudo-random and not known to the subjects. The ordering of the test sequences was also pseudo-random. The TM5 and perceptually coded scenes were displayed for their duration (around five seconds each), and this process was repeated two times. Subjects were then given time to vote on the comparative quality of the sequences (A and B), using the continuous quality comparison scale shown in
Figure 8.7: Timing of stimulus presentation during stimulus-comparison subjective testing.

Figure 8.8. A training set, which used the sequence claire coded at various bit-rates, was used to familiarise the subjects with the testing procedure. However the results of this training set were not used in the final analysis. The duration of a complete test session was around 15 minutes.

The raw subjective test data was averaged across all subjects. The results which show the comparative subjective quality are plotted in Figure 8.9. For both sequences and for both perceptual bit-rates which were tested, the perceptually coded sequences were rated as having higher subjective quality than the TM5 coded sequences of the same bit-rate. The TM5 bit-rate at which its subjective quality equaled the perceptually coded sequences is represented by the point where the comparative quality equals zero. For the Miss America sequence perceptually coded at 250 kbit/sec, the equivalent quality TM5 coded scene was approximately 275 kbit/sec. This represents a 10% increase in bit-rate for the same perceptual quality. At the higher bit-rate (300 kbit/sec), equivalent quality was obtained at a TM5 bit-rate of approximately 375 kbit/sec, which indicates a bit-rate increase of 25%. The improvements were smaller, but still noticeable, for the football sequence (5% and 3% respectively for the lower and higher bit-rate sequences). These results suggest that the improvements in subjective quality obtained by using the perceptual coder are dependent upon the sequences and the
Figure 8.8: Comparison scale used for comparing the qualities of the TM5 and perceptually coded sequences.

Additional subjective experiments, using a larger range of sequences and bit-rates, could be performed in the future to further quantify this relationship.

A consistent and positive improvement in subjective quality was achieved by including perceptual factors in the MPEG coding algorithm. This underlines the value of perceptual coding. The changes to the MPEG TM5 encoder which were performed in this example were quite simple. They consisted of a modification of the quantiser to include the different spatial masking strengths in edge and texture regions, and the inclusion of the IM attention model. Although these modifications were relatively simple, the results achieved were positive, and support the development of more complex perceptual coders. Such coders would be able to take advantage of many other aspects of the encoding process, as discussed in Section 8.1.
8.3 Performance of the Perceptually-Based MPEG Quantiser

Figure 8.9: Comparative subjective quality between the TM5 and perceptually coded sequences (a) Miss America and (b) football. Positive values indicate that the perceptually coded sequence was rated with higher quality, while negative values indicate that the TM5 coded sequence was rated higher.
Chapter 9

Discussion and Conclusions

The rapid increase in the use of compressed digital images has created a strong need for the development of objective picture quality metrics. This thesis has presented a new method for assessing the quality of compressed natural images, based on the operation of the HVS. The early vision model improves on previous techniques by explicitly taking into account human response to complex natural images in the development and tuning of the model. A model of visual attention was then used to take into account the distortion’s location with respect to the scene. The combination of these models has provided a significant increase in the accuracy of the quality metric, as verified by subjective quality experiments. Vision models such as the one presented in this thesis can be used in many areas of image processing. As an example, an MPEG encoder was implemented which used components of the vision model to control the spatially-adaptive quantisation process. The improvements in subjective picture quality achieved by using the perceptual model in the coding algorithm were verified through subjective experiments.

Section 9.1 provides a review of the thesis, focusing on the significant contributions which were made and the results which were achieved. In Section 9.2, areas
in which the models can be improved are identified and possible alternatives are discussed. To conclude in Section 9.3, some application areas which may benefit from the use of vision models such as the one presented in this thesis are discussed.

9.1 Discussion of Achievements

The primary focus of this thesis was the development of an accurate, robust, and efficient technique for assessing the quality of compressed natural pictures. Current techniques for assessing picture quality were reviewed and found to be inadequate. However, objective metrics based on models of early human vision were identified as a promising and intuitive approach. Chapter 2 provided a review of the operation of the HVS. The first sections of this chapter discussed the physiological structure and psychophysical response of early visual processes, focusing on issues relevant to vision models for complex natural images. Previous vision models used for picture quality assessment have not included any higher level or cognitive effects, and assume that all parts of the scene are in foveal view. However, it is known that the location of distortions relative to the scene has a strong influence on overall picture quality. For these reasons, a review of visual attention and eye movements was performed in Section 2.3. An important finding was that the eye movements of different subjects are strongly correlated when viewing typical natural scenes, and that these ROIs usually only constitute a small portion of the scene. A number of bottom-up and top-down factors which affect visual attention and eye movements were identified.

A review of common compression algorithms and the artifacts which they produce was provided in Chapter 3. This chapter also identified the reasons why a picture quality metric is necessary, and listed the desirable features of a quality metric. These include accuracy, speed, robustness, repeatability, multi-dimensional
output formats, and simplicity. Current methods for subjective testing were reviewed, and many inherent problems associated with subjective tests were identified. A review of state-of-the-art objective quality metrics also identified many deficiencies in these techniques, which have slowed their widespread usage by the image processing community. However, promising results have been achieved by fidelity and quality metrics based on HVS models. This approach was identified as being most likely to provide a suitable picture quality metric, and also has application to many other areas of image processing.

In Chapter 4 a critical review of current HVS-based picture quality metrics was performed. Multi-channel HVS models were identified as the best technique for predicting human response to a variety of psychophysical stimuli. However, two key problems of these models were identified when they are used to predict the quality of compressed natural images. Firstly, the model parameters and the structure of the model’s components need to be chosen so that they are appropriate for use with complex natural images. Many previous models have been calibrated using psychophysical data from tests which used very simple and unnatural stimuli. However, the non-linear response of many visual processes results in a reduction in accuracy when these models are used with complex natural images; hence model calibration using more complex, naturalistic stimuli is important for HVS-based quality metrics. Secondly, current HVS-based quality metrics are designed primarily to determine the fidelity of the scene, assuming all areas to be in foveal view. A simple summation of the fidelity map is used to predict picture quality. This fails to take into consideration any higher level or cognitive issues, which are known to have a strong influence on subjects’ perception of overall picture quality. In particular, the location of a distortion with respect to the scene is not taken into account by current vision models.

By taking into consideration the weaknesses of previous early vision models used for quality assessment, a new early vision model was developed and was presented
in Chapter 5. This model is similar in structure to other early vision models, and is based on the response of neurons in the primary visual cortex. However, it has been tuned and structured specifically for complex natural images. To test the accuracy of the model, two processes were used: visual verification of the PDMs, and correlation of PQRs with subjective quality ratings obtained using a Rec. 500 compliant DSCQS test. These tests indicated that the model could accurately predict the fidelity of compressed images, but that there was still room for improvement in the model’s ability to predict picture quality. This confirmed that a simple Minkowski summation of the visible errors did not capture all of the effects known to be involved in human perception of picture quality.

The significance that the location of the distortions has on overall picture quality, combined with the similarity in eye movements of different subjects while viewing natural scenes, motivated the development of the IM model of visual attention in Chapter 6. The IM technique is able to detect ROIs in a scene without requiring any supervision or user input. It is based on five factors which are known to influence attention, but has been designed in a flexible manner, so new factors can easily be included. Eye movement experiments were performed to verify the accuracy of the model, and a strong correlation with subject’s fixations was found. The model was further extended from still images to video by the inclusion of an object motion parameter.

A new quality metric was proposed in Chapter 7. This metric used the output of the IM model to weight the visible distortions detected by the new early vision model, prior to a summation of errors. A comparison of this model’s output with subjective MOS data showed that a significant improvement in correlation was achieved by including the IM model in the quality metric. This is the first HVS-based quality metric to take into account the location of distortions when determining overall picture quality.

The vision models presented in Chapters 5 and 6 can be used in many other
areas of image processing. To demonstrate this, the models were used to control
the adaptive quantisation process in an MPEG encoder in Chapter 8. The com-
ponents of the perceptual model used to perform this were a simplified masking
model and the IM attention model. Subjective testing was used to confirm the
improved subjective quality which can be achieved by using perceptual models
in a compression algorithm, in comparison to conventional techniques.

9.2 Extensions to the Model

Although the subjective testing has confirmed the performance of the vision mod-
els presented in this thesis, the area of vision modeling is still maturing. There
are many areas which still require further research. This section discusses some
of the improvements and extensions which may be made to the models presented
in this thesis.

9.2.1 The Early Vision Model

Two main areas exist for extending the early vision model.

- *Extension to colour*. The model presented in Chapter 5 is designed for
  monochrome images. Extension of the model to include colour information
  should be relatively straightforward. This would involve the inclusion of
two chrominance channels. These channels would have a similar structure
to the luminance channel, but would be calibrated based on human chroma
sensitivity.

- *Extension to video*. This requires the inclusion of temporal as well as spa-
tial channels. Two temporal channels are generally considered sufficient to
model human visual operation: one sustained (low-pass) and one transient
9.2 Extensions to the Model

(band-pass) [104]. Eye movements pose a problem to temporal vision models, since the image velocity and retinal velocity will differ if eye movements are present. Models which fail to take into account viewer eye movements will predict lower acuity in moving areas of the scene (since temporal frequency is high); however acuity will be high if the object is being tracked. Westen et al. [278] have proposed a solution to the problem, by assuming that any steadily moving object in a scene is a candidate for being tracked. Therefore, all such steadily moving objects are compensated for prior to spatio-temporal filtering. This seems to be a valid solution; however it depends upon the ability of the algorithm to detect all trackable, moving areas in the scene. A further discussion on extending the early vision model to video is contained in [187].

Other issues, such as the optimal choice of filterbank, the number of filters, and the extent of masking, are still being researched and debated. However, it appears that the gains to be achieved by adding further complications to the current early vision models are small, in comparison to the improvements which can be achieved through the use of higher level perceptual and cognitive elements.

9.2.2 The Importance Map Model

The development of attention models for complex natural images is a relatively new area of research. The IM model was designed to be computationally inexpensive, while still being able to capture the factors which influence human visual attention. Several areas of further work still remain, and some of these are listed below.

- More complex and accurate algorithms for detecting attentional features could be developed. The current techniques are calibrated empirically, due
to a lack of subjective data. Further experiments could be devised to provide a more solid means for performing model design and calibration.

- A more complicated and accurate segmentation algorithm could be used, which is able to extract regions which correspond to real objects in the scene. This would be facilitated by the use of colour and/or motion information.

- More features, such as colour, face/skin detection [36, 110, 221], or top-down influence, can be included in the model. The flexible structure of the IM model enables the easy incorporation of additional factors.

- The relative weighting of the features can be improved from the equal weighting which is currently implemented. This process should be scene-dependent, since the relative weighting of the factors will vary depending on the scene content.

### 9.2.3 The Quality Metric

Several open issues exist with the development of a quality metric. The most significant of these are listed below.

- A more complicated technique for combining the early vision and attention models could be included. This may require post-processing of the IMs, to increase the extent of the regions of high importance. The IMs may provide feedback to processes in the early vision model, such as the CSF and masking.

- The quality metric presented in this thesis considers the relative location of distortions in the calculation of overall picture quality. However, it is known that the structure of the errors can also influence the subjective quality of a picture. In general, artifacts which have a regular structure
and which add content to the scene will have a stronger effect on overall picture quality than those which are more random in structure, or which remove scene content [132, 289].

- Many other cognitive and memory issues associated with picture quality evaluation remain to be explored. These effects are particularly apparent when video quality is considered. Some of these effects have recently started to be researched and models have been proposed to account for their effects [50, 103, 234]. Issues related to naturalness and image quality are also significant [49, 125] and warrant further investigation.

9.2.4 Perceptual Coding

Perceptual coding is currently an area of active research, and many ways of including vision models in a coder are being developed. The MPEG quantisation controller presented in Chapter 8 demonstrated one way that this can be performed. Further extensions which would be possible within the scope of MPEG would be the development of a more complex masking model, a scene-adaptive QM optimisation, and pre- and post-processing of the scenes. The new MPEG-4 standard will be particularly amenable to inclusion of an attention model such as the IM model. This is because it is object-based, and allows different objects to be coded with different spatial and temporal resolutions. As discussed in Section 8.1, virtually any type of compression algorithm can benefit from the inclusion of a perceptual model.

9.3 Potential Applications

Vision models such as the ones presented in this thesis can be used in many areas of image processing. Some of the most obvious applications which could benefit
from the use of vision models are listed below.

- **Compression.** Perceptual coding is already an area of active research, and perceptual models can be useful in many aspects of the coding process. Both early vision and attention models can be useful in this area.

- **Image and video databases.** The search for particular objects in large databases can be very computationally expensive. The use of an attention model, to extract the few ROIs in each scene in the database a priori, would result in a major reduction in the computational expense of search. An early vision model may also be useful in the search process, to compare the object being searched for with the contents of the database.

- **Digital watermarking.** An early vision model can be used to ensure that the introduced watermark remains below the threshold of detection. An attention model could be used in one of two ways: either to allow stronger watermarking in areas which are not ROIs (since the watermarks are less likely to be objectionable in these areas), or to confine the watermark to the ROIs (since subsequent degradations of the picture are less likely to occur in the ROIs, hence making the watermarking process more robust).

- **Machine vision.** Real-time, computationally expensive tasks such as robot navigation can benefit from the use of an attention model. This enables the most important areas of the scene to be extracted and processed, at the expense of less important regions. Real-time early vision models may also be useful in automated inspection or detection tasks (e.g. [21]), where the operation of a human observer is being simulated.

These examples provide an overview of areas which can benefit from the use of perceptual vision models; many more are likely to exist or be developed in the future. Any application requiring variable-resolution processing of images may
benefit from the use of attention models. The numerous and widespread nature of these applications warrants the continued investigation and development of models of human vision for image processing applications.
Appendix A

Test Images and their Importance Maps
Figure A.1: Test image *announcer*. (a) Original image, and (b) Importance Map.
Figure A.2: Test image beach. (a) Original image, and (b) Importance Map.
Figure A.3: Test image boats. (a) Original image, and (b) Importance Map.
Figure A.4: Test image claire. (a) Original image, and (b) Importance Map.
Figure A.5: Test image football. (a) Original image, and (b) Importance Map.
Figure A.6: Test image *island*. (a) Original image, and (b) Importance Map.
Figure A.7: Test image lena. (a) Original image, and (b) Importance Map.
Figure A.8: Test image *lighthouse*. (a) Original image, and (b) Importance Map.
Figure A.9: Test image Miss America. (a) Original image, and (b) Importance Map.
Figure A.10: Test image *pens*. (a) Original image, and (b) Importance Map.
Figure A.11: Test image *splash*. (a) Original image, and (b) Importance Map.
Appendix B

Individual Quality Ratings for DSCQs Test
<table>
<thead>
<tr>
<th>Scene, Coder, Subject #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann, JPEG, 0.92</td>
<td>5.45</td>
<td>5.05</td>
<td>5.35</td>
<td>5.02</td>
<td>4.97</td>
<td>5.67</td>
<td>5.00</td>
<td>3.95</td>
<td>4.22</td>
</tr>
<tr>
<td>Ann, JPEG, 0.51</td>
<td>4.60</td>
<td>4.22</td>
<td>4.22</td>
<td>4.15</td>
<td>5.00</td>
<td>5.05</td>
<td>4.82</td>
<td>4.62</td>
<td>4.92</td>
</tr>
<tr>
<td>Ann, JPEG, 0.38</td>
<td>4.65</td>
<td>4.17</td>
<td>3.35</td>
<td>4.25</td>
<td>5.00</td>
<td>4.22</td>
<td>4.80</td>
<td>3.42</td>
<td>4.40</td>
</tr>
<tr>
<td>Ann, JPEG, 0.30</td>
<td>4.00</td>
<td>4.20</td>
<td>2.60</td>
<td>3.62</td>
<td>4.20</td>
<td>4.12</td>
<td>3.20</td>
<td>2.87</td>
<td>4.50</td>
</tr>
<tr>
<td>Ann, w’let, 1.63</td>
<td>4.65</td>
<td>5.75</td>
<td>4.15</td>
<td>4.20</td>
<td>5.00</td>
<td>5.00</td>
<td>5.45</td>
<td>5.00</td>
<td>4.87</td>
</tr>
<tr>
<td>Ann, w’let, 0.98</td>
<td>4.82</td>
<td>4.22</td>
<td>4.20</td>
<td>4.17</td>
<td>5.00</td>
<td>5.00</td>
<td>5.57</td>
<td>4.60</td>
<td>4.90</td>
</tr>
<tr>
<td>Ann, w’let, 0.49</td>
<td>4.75</td>
<td>4.22</td>
<td>4.45</td>
<td>5.17</td>
<td>4.50</td>
<td>4.15</td>
<td>4.60</td>
<td>5.65</td>
<td>5.00</td>
</tr>
<tr>
<td>Ann, w’let, 0.26</td>
<td>3.47</td>
<td>3.35</td>
<td>3.32</td>
<td>4.22</td>
<td>3.25</td>
<td>4.20</td>
<td>3.70</td>
<td>4.50</td>
<td>4.20</td>
</tr>
<tr>
<td>Ann, v-JPEG, 0.45</td>
<td>5.00</td>
<td>5.60</td>
<td>4.75</td>
<td>5.00</td>
<td>5.00</td>
<td>5.05</td>
<td>4.17</td>
<td>3.97</td>
<td>5.55</td>
</tr>
<tr>
<td>Ann, v-JPEG, 0.33</td>
<td>5.50</td>
<td>4.25</td>
<td>3.35</td>
<td>4.10</td>
<td>3.32</td>
<td>5.00</td>
<td>5.27</td>
<td>5.00</td>
<td>5.42</td>
</tr>
<tr>
<td>Ann, v-w’let, 0.30</td>
<td>4.95</td>
<td>4.12</td>
<td>3.57</td>
<td>3.90</td>
<td>4.05</td>
<td>4.07</td>
<td>3.67</td>
<td>4.50</td>
<td>4.97</td>
</tr>
<tr>
<td>Bea, JPEG, 1.02</td>
<td>5.30</td>
<td>4.15</td>
<td>4.30</td>
<td>5.72</td>
<td>5.00</td>
<td>4.20</td>
<td>4.47</td>
<td>5.40</td>
<td>5.25</td>
</tr>
<tr>
<td>Bea, JPEG, 0.55</td>
<td>4.75</td>
<td>4.15</td>
<td>4.75</td>
<td>5.20</td>
<td>4.47</td>
<td>5.02</td>
<td>3.75</td>
<td>4.42</td>
<td>5.57</td>
</tr>
<tr>
<td>Bea, JPEG, 0.38</td>
<td>5.42</td>
<td>3.52</td>
<td>3.07</td>
<td>4.45</td>
<td>4.50</td>
<td>4.97</td>
<td>3.62</td>
<td>4.37</td>
<td>5.00</td>
</tr>
<tr>
<td>Bea, JPEG, 0.28</td>
<td>3.75</td>
<td>3.40</td>
<td>2.60</td>
<td>3.85</td>
<td>2.97</td>
<td>5.00</td>
<td>3.35</td>
<td>1.80</td>
<td>4.47</td>
</tr>
<tr>
<td>Bea, w’let, 1.02</td>
<td>4.85</td>
<td>4.30</td>
<td>4.50</td>
<td>4.42</td>
<td>5.00</td>
<td>5.80</td>
<td>5.47</td>
<td>5.00</td>
<td>5.00</td>
</tr>
<tr>
<td>Bea, w’let, 0.61</td>
<td>4.95</td>
<td>4.12</td>
<td>4.05</td>
<td>4.20</td>
<td>4.55</td>
<td>4.02</td>
<td>4.70</td>
<td>4.27</td>
<td>4.97</td>
</tr>
<tr>
<td>Bea, w’let, 0.31</td>
<td>3.70</td>
<td>3.45</td>
<td>3.22</td>
<td>3.90</td>
<td>1.90</td>
<td>4.17</td>
<td>1.90</td>
<td>2.82</td>
<td>5.00</td>
</tr>
<tr>
<td>Bea, w’let, 0.16</td>
<td>3.80</td>
<td>2.52</td>
<td>2.30</td>
<td>3.52</td>
<td>2.32</td>
<td>4.22</td>
<td>1.97</td>
<td>1.22</td>
<td>4.42</td>
</tr>
<tr>
<td>Bea, v-JPEG, 0.41</td>
<td>5.32</td>
<td>4.05</td>
<td>3.85</td>
<td>4.35</td>
<td>5.00</td>
<td>5.00</td>
<td>5.35</td>
<td>5.95</td>
<td>4.70</td>
</tr>
<tr>
<td>Bea, v-JPEG, 0.35</td>
<td>5.00</td>
<td>3.80</td>
<td>3.90</td>
<td>4.20</td>
<td>4.72</td>
<td>4.92</td>
<td>4.70</td>
<td>4.30</td>
<td>5.00</td>
</tr>
<tr>
<td>Bea, v-w’let, 0.27</td>
<td>5.00</td>
<td>4.22</td>
<td>4.10</td>
<td>4.67</td>
<td>4.55</td>
<td>4.25</td>
<td>3.92</td>
<td>3.25</td>
<td>5.05</td>
</tr>
<tr>
<td>Lig, JPEG, 1.41</td>
<td>4.70</td>
<td>5.80</td>
<td>4.15</td>
<td>4.50</td>
<td>4.97</td>
<td>5.00</td>
<td>5.00</td>
<td>4.80</td>
<td>5.47</td>
</tr>
<tr>
<td>Lig, JPEG, 0.91</td>
<td>4.65</td>
<td>5.82</td>
<td>4.05</td>
<td>5.72</td>
<td>5.00</td>
<td>4.30</td>
<td>4.17</td>
<td>5.25</td>
<td>4.77</td>
</tr>
<tr>
<td>Lig, JPEG, 0.69</td>
<td>4.52</td>
<td>4.17</td>
<td>3.70</td>
<td>4.50</td>
<td>5.00</td>
<td>5.00</td>
<td>4.40</td>
<td>3.45</td>
<td>3.65</td>
</tr>
</tbody>
</table>

continued on next page
### Table B.1: MOS for subjects 1–9 for each coded image, following normalisation with respect to the original image. Values greater than 5.00 indicate that the coded scene was rated as having higher quality than the original scene. Ann = Announcer, Bea = Beach, Lig = Lighthouse, Spl = Splash, w’let = wavelet, v-JPEG = variable-quality JPEG, v-w’let = variable-quality wavelet.

<table>
<thead>
<tr>
<th>Scene, Coder, Bit-rate (bit/pixel)</th>
<th>Subject #</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Lig, JPEG, 0.46</td>
<td>4.20</td>
</tr>
<tr>
<td>Lig, w’let, 1.02</td>
<td>5.02</td>
</tr>
<tr>
<td>Lig, w’let, 0.41</td>
<td>4.65</td>
</tr>
<tr>
<td>Lig, w’let, 0.20</td>
<td>3.10</td>
</tr>
<tr>
<td>Lig, v-JPEG, 0.76</td>
<td>4.97</td>
</tr>
<tr>
<td>Lig, v-JPEG, 0.56</td>
<td>5.05</td>
</tr>
<tr>
<td>Lig, v-w’let, 0.35</td>
<td>5.30</td>
</tr>
<tr>
<td>Spl, JPEG, 0.66</td>
<td>5.40</td>
</tr>
<tr>
<td>Spl, JPEG, 0.44</td>
<td>4.77</td>
</tr>
<tr>
<td>Spl, JPEG, 0.29</td>
<td>5.00</td>
</tr>
<tr>
<td>Spl, JPEG, 0.23</td>
<td>4.22</td>
</tr>
<tr>
<td>Spl, w’let, 0.70</td>
<td>5.50</td>
</tr>
<tr>
<td>Spl, w’let, 0.42</td>
<td>4.55</td>
</tr>
<tr>
<td>Spl, w’let, 0.28</td>
<td>4.82</td>
</tr>
<tr>
<td>Spl, w’let, 0.17</td>
<td>4.20</td>
</tr>
<tr>
<td>Spl, v-JPEG, 0.41</td>
<td>4.95</td>
</tr>
<tr>
<td>Spl, v-JPEG, 0.31</td>
<td>5.12</td>
</tr>
<tr>
<td>Spl, v-w’let, 0.27</td>
<td>5.05</td>
</tr>
</tbody>
</table>
### Individual Quality Ratings for DSCQS Test

<table>
<thead>
<tr>
<th>Scene, Coder, Bit-rate (bit/pixel)</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>1</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann, JPEG, 0.92</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>4.85</td>
<td>5.07</td>
<td>5.00</td>
<td>5.05</td>
<td>5.00</td>
<td>5.15</td>
</tr>
<tr>
<td>Ann, JPEG, 0.51</td>
<td>4.80</td>
<td>5.00</td>
<td>5.75</td>
<td>4.57</td>
<td>4.22</td>
<td>5.40</td>
<td>5.00</td>
<td>4.00</td>
<td>4.20</td>
</tr>
<tr>
<td>Ann, JPEG, 0.38</td>
<td>4.80</td>
<td>4.15</td>
<td>3.30</td>
<td>4.55</td>
<td>4.60</td>
<td>5.00</td>
<td>4.45</td>
<td>3.37</td>
<td>4.20</td>
</tr>
<tr>
<td>Ann, JPEG, 0.30</td>
<td>4.45</td>
<td>4.30</td>
<td>2.25</td>
<td>4.17</td>
<td>4.20</td>
<td>3.77</td>
<td>3.50</td>
<td>3.60</td>
<td>2.57</td>
</tr>
<tr>
<td>Ann, w’let, 1.63</td>
<td>4.95</td>
<td>5.00</td>
<td>4.10</td>
<td>5.12</td>
<td>5.00</td>
<td>5.50</td>
<td>5.35</td>
<td>5.30</td>
<td>5.10</td>
</tr>
<tr>
<td>Ann, w’let, 0.98</td>
<td>5.10</td>
<td>5.85</td>
<td>4.90</td>
<td>5.00</td>
<td>5.25</td>
<td>4.67</td>
<td>5.00</td>
<td>4.80</td>
<td>4.75</td>
</tr>
<tr>
<td>Ann, w’let, 0.49</td>
<td>5.07</td>
<td>5.00</td>
<td>4.45</td>
<td>5.07</td>
<td>4.50</td>
<td>3.82</td>
<td>4.60</td>
<td>4.40</td>
<td>4.42</td>
</tr>
<tr>
<td>Ann, w’let, 0.26</td>
<td>4.75</td>
<td>4.15</td>
<td>4.00</td>
<td>4.85</td>
<td>3.35</td>
<td>3.75</td>
<td>4.80</td>
<td>3.22</td>
<td>2.22</td>
</tr>
<tr>
<td>Ann, v-JPEG, 0.45</td>
<td>4.77</td>
<td>3.50</td>
<td>3.90</td>
<td>4.70</td>
<td>3.97</td>
<td>4.65</td>
<td>4.65</td>
<td>3.72</td>
<td>4.20</td>
</tr>
<tr>
<td>Ann, v-JPEG, 0.33</td>
<td>4.55</td>
<td>3.35</td>
<td>3.50</td>
<td>4.27</td>
<td>3.97</td>
<td>4.75</td>
<td>5.00</td>
<td>2.82</td>
<td>4.00</td>
</tr>
<tr>
<td>Ann, v-w’let, 0.30</td>
<td>4.85</td>
<td>4.20</td>
<td>3.92</td>
<td>4.85</td>
<td>4.07</td>
<td>4.45</td>
<td>5.00</td>
<td>3.60</td>
<td>4.55</td>
</tr>
<tr>
<td>Bea, JPEG, 1.02</td>
<td>4.85</td>
<td>4.15</td>
<td>4.97</td>
<td>5.00</td>
<td>4.85</td>
<td>5.00</td>
<td>5.00</td>
<td>4.75</td>
<td>4.95</td>
</tr>
<tr>
<td>Bea, JPEG, 0.55</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>4.92</td>
<td>5.00</td>
<td>5.47</td>
<td>4.35</td>
<td>5.00</td>
<td>5.00</td>
</tr>
<tr>
<td>Bea, JPEG, 0.38</td>
<td>4.62</td>
<td>4.10</td>
<td>3.57</td>
<td>4.75</td>
<td>4.47</td>
<td>5.00</td>
<td>4.42</td>
<td>4.20</td>
<td>3.15</td>
</tr>
<tr>
<td>Bea, JPEG, 0.28</td>
<td>4.42</td>
<td>2.47</td>
<td>2.57</td>
<td>4.85</td>
<td>3.75</td>
<td>5.00</td>
<td>3.45</td>
<td>3.60</td>
<td>2.07</td>
</tr>
<tr>
<td>Bea, w’let, 1.02</td>
<td>4.87</td>
<td>4.95</td>
<td>5.00</td>
<td>5.05</td>
<td>4.77</td>
<td>5.00</td>
<td>4.70</td>
<td>4.75</td>
<td>4.47</td>
</tr>
<tr>
<td>Bea, w’let, 0.61</td>
<td>4.75</td>
<td>4.10</td>
<td>4.35</td>
<td>4.90</td>
<td>4.45</td>
<td>4.55</td>
<td>4.75</td>
<td>4.42</td>
<td>4.40</td>
</tr>
<tr>
<td>Bea, w’let, 0.31</td>
<td>4.17</td>
<td>4.15</td>
<td>2.60</td>
<td>4.90</td>
<td>4.05</td>
<td>4.82</td>
<td>3.82</td>
<td>3.60</td>
<td>1.80</td>
</tr>
<tr>
<td>Bea, w’let, 0.16</td>
<td>3.82</td>
<td>1.80</td>
<td>2.05</td>
<td>4.20</td>
<td>3.00</td>
<td>3.67</td>
<td>3.17</td>
<td>2.60</td>
<td>1.60</td>
</tr>
<tr>
<td>Bea, v-JPEG, 0.41</td>
<td>4.90</td>
<td>5.00</td>
<td>5.02</td>
<td>5.10</td>
<td>4.60</td>
<td>4.75</td>
<td>5.20</td>
<td>4.12</td>
<td>4.77</td>
</tr>
<tr>
<td>Bea, v-JPEG, 0.35</td>
<td>4.85</td>
<td>4.15</td>
<td>5.00</td>
<td>4.80</td>
<td>5.20</td>
<td>4.67</td>
<td>4.25</td>
<td>4.52</td>
<td>5.00</td>
</tr>
<tr>
<td>Bea, v-w’let, 0.27</td>
<td>4.82</td>
<td>5.00</td>
<td>5.02</td>
<td>4.90</td>
<td>5.00</td>
<td>4.15</td>
<td>4.55</td>
<td>4.40</td>
<td>4.85</td>
</tr>
<tr>
<td>Lig, JPEG, 1.41</td>
<td>5.00</td>
<td>5.00</td>
<td>4.60</td>
<td>4.87</td>
<td>5.17</td>
<td>5.07</td>
<td>5.00</td>
<td>5.02</td>
<td>4.95</td>
</tr>
<tr>
<td>Lig, JPEG, 0.91</td>
<td>5.00</td>
<td>4.95</td>
<td>4.85</td>
<td>4.75</td>
<td>4.50</td>
<td>5.85</td>
<td>5.00</td>
<td>4.20</td>
<td>4.20</td>
</tr>
<tr>
<td>Lig, JPEG, 0.69</td>
<td>5.00</td>
<td>5.00</td>
<td>3.80</td>
<td>4.80</td>
<td>3.90</td>
<td>4.55</td>
<td>3.82</td>
<td>4.45</td>
<td>4.22</td>
</tr>
</tbody>
</table>

continued on next page
<table>
<thead>
<tr>
<th>Scene, Coder, Bit-rate (bit/pixel)</th>
<th>Subject #</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Lig, JPEG, 0.46</td>
<td>4.85</td>
</tr>
<tr>
<td>Lig, w’let, 1.02</td>
<td>5.00</td>
</tr>
<tr>
<td>Lig, w’let, 0.41</td>
<td>4.57</td>
</tr>
<tr>
<td>Lig, w’let, 0.20</td>
<td>3.60</td>
</tr>
<tr>
<td>Lig, v-JPEG, 0.76</td>
<td>4.97</td>
</tr>
<tr>
<td>Lig, v-JPEG, 0.56</td>
<td>5.00</td>
</tr>
<tr>
<td>Lig, v-w’let, 0.35</td>
<td>4.57</td>
</tr>
<tr>
<td>Spl, JPEG, 0.66</td>
<td>5.00</td>
</tr>
<tr>
<td>Spl, JPEG, 0.44</td>
<td>5.05</td>
</tr>
<tr>
<td>Spl, JPEG, 0.29</td>
<td>4.85</td>
</tr>
<tr>
<td>Spl, JPEG, 0.23</td>
<td>4.72</td>
</tr>
<tr>
<td>Spl, w’let, 0.70</td>
<td>5.00</td>
</tr>
<tr>
<td>Spl, w’let, 0.42</td>
<td>4.92</td>
</tr>
<tr>
<td>Spl, w’let, 0.28</td>
<td>4.90</td>
</tr>
<tr>
<td>Spl, w’let, 0.17</td>
<td>4.67</td>
</tr>
<tr>
<td>Spl, v-JPEG, 0.41</td>
<td>5.00</td>
</tr>
<tr>
<td>Spl, v-JPEG, 0.31</td>
<td>4.72</td>
</tr>
<tr>
<td>Spl, v-w’let, 0.27</td>
<td>4.90</td>
</tr>
</tbody>
</table>

Table B.2: MOS for subjects 10–18 for each coded image, following normalisation with respect to the original image. Values greater than 5.00 indicate that the coded scene was rated as having higher quality than the original scene. Ann = Announcer, Bea = Beach, Lig = Lighthouse, Spl = Splash, w’let = wavelet, v-JPEG = variable-quality JPEG, v-w’let = variable-quality wavelet.
Appendix C

Fixation Points Across All Subjects
Figure C.1: Fixations across all subjects for image announcer.

Figure C.2: Fixations across all subjects for image beach.
Figure C.3: Fixations across all subjects for image *boats*.

Figure C.4: Fixations across all subjects for image *claire*.
Figure C.5: Fixations across all subjects for image *football*.

Figure C.6: Fixations across all subjects for image *island*.
**Figure C.7:** Fixations across all subjects for image *lena*.

**Figure C.8:** Fixations across all subjects for image *light*.
Figure C.9: Fixations across all subjects for image Miss America.

Figure C.10: Fixations across all subjects for image pens.
Figure C.11: Fixations across all subjects for image splash.
Bibliography


[71] M. P. Eckert. Lossy compression using wavelets, block DCT, and lapped orthogonal transforms optimized with a perceptual model. In *Proceedings*


[79] EPWIC-1 (Embedded Predictive Wavelet Image Coder), available from ftp.cis.upenn.edu/pub/butch/EPWIC.

243


[120] Independent JPEG Group codec, available from ftp.uu.net/graphics/jpeg.


250


