Visual search in natural scenes with and without guidance of fixations

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2011

By
Matthew Simon Mould
School of Electrical and Electronic Engineering
Contents

Abstract 14

Declaration 16

Copyright 17

Acknowledgements 19

1 Introduction 20
   1.1 Basic physiology of the eye ........................................ 21
   1.2 Eye movements ...................................................... 23
   1.3 Eye-tracking ......................................................... 24
   1.4 Classifying eye movements ......................................... 25
   1.5 Visual search ....................................................... 27
   1.6 Predicting human eye movements ................................. 29
   1.7 Steering visual attention .......................................... 31
      1.7.1 Adjustment of local contrast ................................. 31
      1.7.2 Adjusting texture contrast ................................... 33
      1.7.3 Guiding attention by geometry .............................. 33
      1.7.4 Guiding attention by subtle flicker ......................... 34
      1.7.5 Success in steering attention may depend on task ....... 34
1.7.6 Image enhancement used in this project . . . . . . . . . . . . 35
1.8 Colour spaces . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
1.9 Information theory . . . . . . . . . . . . . . . . . . . . . . . . . . 40
  1.9.1 Entropy . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40
  1.9.2 KL divergence . . . . . . . . . . . . . . . . . . . . . . . . . 42
1.10 Thesis outline . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42

2 Kullback-Leibler Divergence 45
  2.1 Methods of estimating KL divergence . . . . . . . . . . . . . . . 46
    2.1.1 Synthetic data set . . . . . . . . . . . . . . . . . . . . . . . 47
    2.1.2 Binning and adding a constant . . . . . . . . . . . . . . . . 49
    2.1.3 Kernel smoothing . . . . . . . . . . . . . . . . . . . . . . . 52
    2.1.4 Leonenko et al.’s KL divergence estimator . . . . . . . . . . 55
  2.2 Testing of Leonenko et al.’s KL divergence estimator against an ana-
    lytical solution . . . . . . . . . . . . . . . . . . . . . . . . . . . . 56
    2.2.1 Analytical expression for KL divergence between two Gauss-
      sian distributions . . . . . . . . . . . . . . . . . . . . . . . . 57
    2.2.2 Use of Leonenko et al. estimator to calculate KL divergence
      between two Gaussian functions . . . . . . . . . . . . . . . 60
    2.2.3 The offset version of the Leonenko et al. estimator . . . . . 60
  2.3 Behaviour of Leonenko et al.’s (2008) KL estimator with data sampled
      from identical distributions . . . . . . . . . . . . . . . . . . . . 64
  2.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65

3 Collection and Analysis of Eye Movement Data 67
  3.1 The CRS Video Eyetracker Toolbox . . . . . . . . . . . . . . . . 67
    3.1.1 Testing with Artificial Eye . . . . . . . . . . . . . . . . . . 68
    3.1.2 Calibration and testing with human observers . . . . . . . . 70
<table>
<thead>
<tr>
<th>Section</th>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>Classification of fixations</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>3.2.1 Data</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>3.2.2 The method due to van der Linde et al.</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>3.2.3 The method due to Vig et al.</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>3.2.4 An objective method for classifying fixations</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>3.2.5 Classification method</td>
<td>77</td>
</tr>
<tr>
<td>3.3</td>
<td>Discussion</td>
<td>89</td>
</tr>
<tr>
<td>4</td>
<td>Searching Images of Natural Scenes</td>
<td>91</td>
</tr>
<tr>
<td>4.1</td>
<td>Method</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>4.1.1 Apparatus</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>4.1.2 Stimuli</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>4.1.3 Observers</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>4.1.4 Experimental procedure</td>
<td>95</td>
</tr>
<tr>
<td>4.2</td>
<td>Results and analysis</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>4.2.1 Luminance profile</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>4.2.2 Chroma profile</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>4.2.3 Luminance entropy profile</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>4.2.4 Luminance contrast profile</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>4.2.5 Edge density profile produced using Gabor filters</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>4.2.6 Edge density profile produced using Canny edge detector</td>
<td>102</td>
</tr>
<tr>
<td></td>
<td>4.2.7 Feature congestion profile</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>4.2.8 Subband entropy profile</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>4.2.9 Variance accounted for by each feature</td>
<td>104</td>
</tr>
<tr>
<td>4.3</td>
<td>Discussion</td>
<td>108</td>
</tr>
<tr>
<td>5</td>
<td>Steering Human Fixations</td>
<td>111</td>
</tr>
<tr>
<td>5.1</td>
<td>Smoothly decreasing de-enhancement</td>
<td>112</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Agreement (%) between fixation classifications by each experimental method and by three experts across 21 sample traces (5271 frames). Expert RA classified all 21 traces; expert EG classified 18 traces (4518 frames); expert GRB classified 3 traces (753 frames). . . . . . . . . . 87
List of Figures

1.1 An image of a festival in Manchester. . . . . . . . . . . . . . . . . . 30
1.2 An image of two zebra. The right hand side of the image has been sub-
ject to unsharp masking with k=0.9. The left hand side is unmodified. 36
2.1 Entropy of a data sets sampled from a Gaussian distribution with mean
0 and standard deviation 1 estimated using the Kozachenko-Leonenko
estimator described by Leonenko et al. (2008). The values shown are
the means of 100 simulations. . . . . . . . . . . . . . . . . . . . . . 46
2.2 KL divergence between two distributions composed of superpositions
of Gaussian functions (see Section 2.1.1 for details) calculated from a
discrete approximation to the underlying probability density function
as the number of bins was varied. . . . . . . . . . . . . . . . . . . . 48
2.3 KL divergence between two data sets sampled from distributions com-
possed of superpositions of Gaussian functions (see Section 2.1.1 for
details) estimated using the method described in Section 2.1.2 as N
and p were varied and r was set to 20. Results shown are the means
taken over 100 runs. . . . . . . . . . . . . . . . . . . . . . . . . . . . 49
2.4 KL divergence between two data sets sampled from distributions composed of superpositions of Gaussian functions (see Section 2.1.1 for details) estimated using the method described in Section 2.1.2 as $N$ and $r$ were varied and $p$ was set to $10^{-6}$. The upper graph shows the full range of the data, and lower graph has a narrower vertical scale to show detail. Results shown are the means taken over 100 runs. . . . . 51

2.5 KL divergence between two data sets sampled from distributions composed of superpositions of Gaussian functions (see Section 2.1.1 for details) estimated using the method described in Section 2.1.3 as $N$ and $p$ were varied and $r$ was fixed to $2^8$. Results shown are the means taken over 100 runs. . . . . . . . . . . . . . . . . . . . . . . . . . . . 53

2.6 KL divergence between two data sets sampled from distributions composed of superpositions of Gaussian functions (see Section 2.1.1 for details) estimated using the method described in Section 2.1.3 as $N$ and $r$ were varied and $p$ was set to $10^{-6}$. Results shown are the means taken over 100 runs. . . . . . . . . . . . . . . . . . . . . . . . . . . . 54

2.7 KL divergence between two distributions composed of superpositions of Gaussian functions (see Section 2.1.1 for details) estimated using Leonenko et al.’s (2008) estimator as $N$ was varied. Results shown are the means taken over 100 runs. . . . . . . . . . . . . . . . . . . . . . . . . . . . 56

2.8 Kullback Leibler divergence between two data sets sampled from distributions $\mathcal{N}([0, 0], [2, 2])$ and $\mathcal{N}([0, 2], [3, 3])$ estimated using Leonenko et al.’s (2008) estimator, and the offset version of the same estimator (explained in Section 2.2.3) as $N$ was varied. The values shown are the means of 100 simulations. . . . . . . . . . . . . . . . . . . . . . . . . . . . 61
2.9 Kullback Leibler Divergence between two data sets sampled from identical distributions ($\mathcal{N}([0, 0], [1, 1])$) estimated using Leonenko et al.'s (2008) estimator, and the offset version of the same estimator. The values shown are the means of 100 simulations. . . . . . . . . . . . . 65

3.1 A pupil being tracked by the CRS eye-tracker. . . . . . . . . . . . . . 68

3.2 A screen shot of the artificial eye being tracked by monitoring the centre of the pupil and the corneal glints. . . . . . . . . . . . . . 69

3.3 The Vernier scale used to adjust the artificial eye. . . . . . . . . . . . 69

3.4 The number of fixations produced by the method due to van der Linde et al. (2009) for various stability and duration thresholds. . . . . . . . . 74

3.5 The number of fixations produced by the method due to Vig et al. (2009) while varying the stability and duration thresholds. . . . . . . . . 76

3.6 Distribution of point-of-gaze speeds. The grey histogram shows the number of local speed maxima exceeding a variable threshold that ranged between the lowest and highest recorded speed maxima. The dashed line shows the number of local speed maxima exceeding threshold under the null distribution, according to which the maxima are uniformly distributed. A red dotted curve marks the gap between the two, but it is hidden by the solid curve, which shows the loess smooth of the same data. Based on 223 trials by one observer viewing one image. . . 79

3.7 Distribution of candidate fixation durations. The grey histogram shows the number of fixation durations and instrumental noise events at very short durations. The solid curve shows the loess smooth of the histogram. Based on the same set of data as in Figure 3.6. . . . . . . . . . . . . 81
3.8 Distribution of optimum speed thresholds \( v_{opt} \) and optimum duration thresholds \( d_{opt} \) over 140 blocks of 31409 trials (light-grey histograms). Also shown is the distribution of \( v_{opt} \) and \( d_{opt} \) for blocks from which a subsample of 21 trials was taken from the 31409 trials for expert classification (dark-grey histograms). 83

3.9 An example of a typical eye movement trace. Vertical and horizontal gaze locations are plotted as a function of time from the beginning of the sample. Fixations classified by the proposed method (upper plot) and by expert EG (lower plot) are indicated by thicker curves. The classification agreement between these two plots was 95%. A two-dimensional representation of these data is shown in Figure 3.10. 85

3.10 A 2 dimensional plot of a the eye movement trace shown in Figure 3.9. Fixations classified by the proposed method (upper plot) and by expert EG (lower plot) are indicated by the red circles. 86

3.11 Effect of added noise. Optimum speed threshold \( v_{opt} \) averaged over all 140 blocks (upper panel) and corresponding number of fixations classified by the proposed method (lower panel) are plotted against added noise \( \sigma \) in degrees of visual angle. 88

4.1 A road sign warning that cycling is not allowed (Herbert, 2010). 92

4.2 A lizard on some grass (Heffner, 2009). 92

4.3 Some examples of the images used in this experiment. The target is present in the top right image and is indicated with the red arrow. 96

4.4 Some example detection profiles. Brightness corresponds to detection performance measured by \( d' \). 98

4.5 The Gabor filter. 102
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6</td>
<td>The proportion of variance of detection performance accounted for by each feature alone. Error bars represent one standard error.</td>
</tr>
<tr>
<td>4.7</td>
<td>The proportion of variance of detection performance accounted by a linear model of fixation density ( F ) and each image feature ( I ). The model had the form ( d' = aF + bI + cFI ), where ( a, b ) and ( c ) are constants which were selected to optimize the fit. The black vertical line shows the proportion of variance of detection performance accounted for by fixation density alone.</td>
</tr>
<tr>
<td>5.1</td>
<td>An image which had been de-enhanced with maximum de-enhancement at the right (upper image) and the original image for comparison (lower image).</td>
</tr>
<tr>
<td>5.2</td>
<td>Images used in de-enhancement experiments.</td>
</tr>
<tr>
<td>5.3</td>
<td>The mean of horizontal fixation locations under each condition (no de-enhancement, left side de-enhanced and right side de-enhanced) for ramp de-enhancement with target-search task.</td>
</tr>
<tr>
<td>5.4</td>
<td>The Kullback-Leibler divergences between the horizontal component of fixations under different conditions for the ramp de-enhancement with target-search task experiment.</td>
</tr>
<tr>
<td>5.5</td>
<td>The difference between the mean of horizontal fixation positions under the condition with no de-enhancement and the conditions with a de-enhancement for each image for ramp de-enhancement with change-detection experiment.</td>
</tr>
<tr>
<td>5.6</td>
<td>The KL divergence between horizontal components of fixations recorded during each half of each trial for each image and condition ramp de-enhancement with change-detection experiment.</td>
</tr>
</tbody>
</table>
5.7 The difference between the mean of horizontal fixation positions under the condition with no de-enhancement and the conditions with a de-enhancement for each image for ramp de-enhancement with forced-choice target-search experiment. ........................................ 126

5.8 The KL divergence between horizontal components of fixations recorded during each half of each trial for each image and condition ramp de-enhancement with forced-choice target-search experiment. ............ 127

5.9 An example image with an enhanced region, indicated by the red arrow (upper image) and the original image for comparison (lower image). . 129

5.10 Images used in local-enhancement experiments. ....................... 130

5.11 The fixation density in annuli around the enhancement for the search task with dynamic local image enhancement experiment (upper figure) and a log-log fit to the same data (lower figure). .................... 134

5.12 The detection performance for each image when the enhancement and target were together or apart for the search task with dynamic local image enhancement. ........................................ 135

5.13 The detection performance for observers attempting to detect the enhancement for each image for the detection of dynamic local enhancement experiment. ........................................ 137

5.14 The fixation density in annuli around the enhancement for the search task with static local image enhancement experiment (upper figure) and a log-log fit to the same data (lower figure). ................... 140

5.15 The detection performance for each image when the enhancement and target were together or apart for the search task with static local image enhancement experiment. ............................ 141
5.16 The detection performance for observers attempting to detect the enhancement for each image for the detection of static local enhancement experiment.
Abstract

From the airport security guard monitoring luggage to the rushed commuter looking for their car keys, visual search is one of the most common requirements of our visual system. Despite its ubiquity, many aspects of visual search remain unaccounted for by computational models. Difficulty arises when trying to account for any internal biases of an observer undertaking a search task or trying to decompose an image of a natural scene into relevant fundamental properties. Previous studies have attempted to understand visual search by using highly simplified stimuli, such as discrete search arrays. Although these studies have been useful, the extent to which the search of discrete search arrays can represent the search of more naturalistic stimuli is subject to debate.

The experiments described in this thesis used as stimuli images of natural scenes and attempted to address two key objectives. The first was to determine which image properties influenced the detectability of a target. Features investigated included chroma, entropy, contrast, edge contrast and luminance. The proportion of variance in detection ability accounted for by each feature was estimated and the features were ranked in order of importance to detection. The second objective was to develop a method for guiding human fixations by modifying image features while observers were engaged in a search task. To this end, images were modified using the image-processing method unsharp masking. To assess the effect of the image modification on fixations, eye movements were monitored using an eye-tracker.
Another subject addressed in the thesis was the classification of fixations from eye movement data. There exists no standard method for achieving this classification. Existing methods have employed thresholds for speed, acceleration, duration and stability of point-of-gaze to classify fixations, but these thresholds have no commonly accepted values. Presented in this thesis is an automatic nonparametric method for classifying fixations, which extracts fixations without requiring any input parameters from the experimenter. The method was tested against independent classifications by three experts.

The accurate estimation of Kullback-Leibler Divergence, an information theoretic quantity which can be used to compare probability distributions, was also addressed in this thesis since the quantity was used to compare fixation distributions. Different methods for the estimation of Kullback-Leibler divergence were tested using artificial data and it was shown than a method for estimating the quantity directly from input data outperformed methods which required binning of data or kernel density estimation to estimate underlying distributions.
Declaration

No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Experiments described in this thesis were approved by the University of Manchester Committee on the Ethics of Research on Human Beings, which operated according to the principles of the Declaration of Helsinki.
Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University
IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectual-property.pdf), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses
Acknowledgements

I would like to thank my supervisor Prof. David Foster for being extremely generous with support and advice over the course of this project, Dr. Kinjiro Amano for his help and useful suggestions and Dr. John Oakley for his stimulating questions and comments. Many thanks to Dr. Iván Marín-Franch for providing code and discussion on the subject of information theory. Thanks to Dr. Rochelle Ackerley, Prof. Graham Barnes, and Dr. Emma Gowen for their help in analyzing the experimental data and Dr. Eleonora Vig for advice. Thanks also to all of the people from the Sensing, Imaging, and Signal Processing Group at the University of Manchester for providing a friendly working environment. Finally, grateful thanks to my parents Keith and Jeanette, sister Laura and everyone else in my family for their encouragement.

This work was supported by Engineering and Physical Sciences Research Council, grant no. EP/F023669/1.
Chapter 1

Introduction

The human visual system gives the impression of a thorough and complete representation of the outside world, but the amount of time it can take to find a set of keys on a cluttered desk reveals that our visual perception is far from perfect. Because peripheral vision is far less detailed than central vision, it is necessary to move the eyes over the desk to maximize the information collected, and sometimes we can look directly at the keys without perceiving them. The importance of visual search both to the vision science community and in general is reflected by the number of publications which study it. A search of the ISI database performed on 5th October 2011 found over 6000 articles with titles containing the phrase “visual search”. Many attempts to model visual search (summarized in Section 1.5) have vastly simplified the situation by modelling search in artificial search arrays rather than search in more complex stimuli such as images of natural scenes. One of the objectives of this thesis was to assess the factors affecting visual search performance in images of natural scenes, a subject which is addressed in Chapter 4.

Another of the objectives was to develop a method to guide human fixations systematically. In Chapter 5, two approaches to this challenge are described. The first approach aimed to steer fixations away from particular image regions. The motivation
behind this approach was an application in image compression. If observers could be persuaded not to look at particular regions, bandwidth dedicated to that region’s representation could be reduced without an observer noticing. This would make possible the selective compression of images. The second approach aimed to steer fixations towards a particular image region in order to subtly aid visual search without excessively altering the image.

Over the course of this project techniques for the analysis of eye movement data were also studied. In Chapter 2 methods for estimating the Kullback-Leibler Divergence, an information theoretic quantity sometimes used to compare scanpaths, were tested and compared and in Chapter 3 problems in the collection and classification of eye movement data were addressed.

1.1 Basic physiology of the eye

In order to understand the factors which influence human visual attention, it is necessary to review briefly how the human visual system sorts the vast amount of data incident upon it into a meaningful signal. Light is focussed onto the retina by both the cornea and the lens (Schwartz, 1994). During eye movement, the lens’ power is continually adjusted by the ciliary muscles to maintain a sharp image on the retina, where visual processing begins.

The retina

The retina is composed of two varieties of photosensitive cell: rods and cones. Rods are very sensitive to light while the cones function better at higher light levels. The cones can be classified as one of three types; S-cones, which respond to short-wavelengths,
M-cones, which respond to medium-wavelengths and L-cones, which respond to long-wavelengths. Each contains a different type of light absorbing pigment, and so responds to light of different wavelengths. Colour perception is achieved by comparing the output from each type of cone. A very simplified outline of colour-opponency is that the difference in output from the L- and M-cones produces a red-green opponent channel, and the difference between the S-cones and the L- and M-cones’ combined output gives a blue-yellow opponent channel (Dacey and Packer, 2003).

The highest level of visual acuity on the retina is achieved at the fovea. It subtends approximately 1.5 mm, corresponding to 5 deg of visual angle (Polyak, 1957), and is composed almost entirely of cones. The number of cones falls off sharply outside of the fovea, and the regions representing peripheral vision are largely populated by rods. Although the rods are very numerous—each eye contains over 100 million (Wandell, 1995)—they do not provide correspondingly high resolving power since the signals from several rods are passed to single inter-neurones; this arrangement provides greater light sensitivity at the cost of high acuity (Wandell, 1995).

**Receptive fields and centre-surround organisation**

The retina is composed of two synaptic layers. The photoreceptor cells pass information to the bipolar-cell layer, and the signals are modified by horizontal cells. The signals from the bipolar-cell layer are transmitted to the ganglion-cell layer, with these signals being modified by amacrine cells (Baccus, 2007).

Each ganglion cell is sensitive to signals from particular photoreceptor cells, and the region to which a ganglion cell signals a response is called its receptive field. Retinal ganglion cells have approximately circular receptive fields with a centre-surround organisation, meaning that light falling on the centre of a cell’s receptive field will produce a different response from that falling on a peripheral region (Kuffler, 1953).
two types of ganglion cells are those which are excited by light falling on the centre of their receptive field, and those which are inhibited by central illumination, known as on-centre and off-centre ganglion cells respectively (Famiglietti Jr. and Kolb, 1976). The arrangement means that if the entire receptive field is illuminated, a fairly small response is obtained, leaving the response curve of the ganglion cell sensitive to spatial contrast rather than uniformly lit scenes. A visual system which is primarily sensitive to contrasts is advantageous because the absolute intensity of light being reflected from a natural scene does not provide much information about the scene itself. Information about the scene will be contained in the image structure and can be obtained by extracting features such as edge contrast and orientation.

1.2 Eye movements

Because of the distribution of photoreceptor cells on the retina, visual acuity decreases rapidly outside of the central visual field (Jacobs, 1979). This can be illustrated by attempting to read this page without moving your eye from the full stop at the end of this sentence. Since visual acuity is so low in the periphery, visual information is collected by moving the eye over a scene.

The eye repositions itself around three times per second using rapid ballistic movements known as saccades. These can approach velocities of over 700 deg s$^{-1}$ (Carpenter, 1988). Between saccades, when the eye is relatively still, the eye is said to be fixating. Even during a fixation the eye continues to make small movements. These fixational eye movements include tremor, a small rapid movement of the eye with an amplitude usually between about 5 and 60 arcsec; also microsaccades, which are short involuntary eye movements with an amplitude between about 1 and 120 arcmin; and drift, which is a slow eye movement with an amplitude between about 1.2 and 30 arcmin (Martinez-Conde et al., 2004).
1.3 Eye-tracking

Although overt visual attention (gaze location) and covert visual attention (internal visual attention) can be deliberately separated by an observer, they are usually very strongly interlinked (Itti and Koch, 2001; Henderson, 2003; Peters et al., 2005). Eye movement data are therefore often used as an indicator of visual attention. Early eye-tracking methods include the scleral search coil method (Robinson, 1963), in which an observer wore a contact lens containing a search coil while sitting in a magnetic field. Movements of the eye would then generate a voltage across the search coil, which could be measured and be used to monitor eye movements. An example of a non-invasive approach is Cornsweet and Crane’s (1973) Dual-Purkinje-Image Tracker. The method involved no direct contact to the eye. Instead, the eye was illuminated with a collimated beam of infra-red light. The light reflected off the front of the cornea, the back of the cornea, the front of the lens and the back of the lens to produce the first, second, third and fourth Purkinje images. The positions of the first and fourth Purkinje images could be measured and their displacement used to monitor the position of the eye. A digital eye-tracking method was proposed by Barbur et al. (1987). In this method, the eye was illuminated with infra-red light and recorded using CCD camera. The position of the pupil’s centre was extracted by digital methods and could be used as an indicator of point of gaze. As cameras and computers have increased in speed, the temporal resolution of eye-trackers has improved (Clarke et al., 2002; Duchowski, 2007)

In this project, eye movements were monitored using a Cambridge Research Systems (CRS) High-Speed Video Eyetracker Toolbox. It records point of gaze (PoG) by monitoring the positions of two glints of light on the corneal surface and the position of the pupil. The features are tracked by the eye-tracker’s image processing software. The eye-tracker runs at a frame rate of 250 Hz and has a nominal accuracy of 0.125
to 0.25 deg. Testing and assessment of the eye-tracker is discussed in more detail in Chapter 3.

1.4 Classifying eye movements

Since little information is collected during saccades (Thilo et al., 2003), and it is the fixations themselves that are of principal interest to the experimenter, it is usual to summarize recordings of the point of gaze (PoG) by extracting just the fixations and discarding the remainder of the data. Unfortunately, classifying fixations in practise can be difficult, and there is no standard method. Part of the difficulty lies in the fact that, as mentioned before, the eye continues to make small movements—including tremor, microsaccades and drift—during a fixation.

Existing methods for classifying fixations include those based on gaze stability, whose complement is sometimes known as dispersion, and on speed (Salvucci and Goldberg, 2000). A stability-based method defines a sequence of eye movements as a fixation if the PoG remains within a circle of given radius (the stability threshold) for a given duration (the duration threshold) (van der Linde et al., 2009); analogously, a speed-based method defines a series of eye movements as a fixation if eye speed remains below a given value (the speed threshold) (Kienzle et al., 2009). The two methods may be combined by incorporating both stability and speed thresholds, sometimes also with an acceleration threshold (Tatler, 2007).

The absence of a standard method for classifying PoG data presents a problem, as does the absence of commonly accepted threshold values for speed, acceleration, duration, and stability. It is known that different choices of algorithms and their input parameters can lead to systematic differences in reported fixations, and markedly different interpretations of gaze data (Shic et al., 2008), as will be shown in Chapter 3. The uncertainty about appropriate parameter values is not surprising given that the
spatiotemporal characteristics of eye movements may vary across observers, stimuli, and tasks (Andrews and Coppola, 1999). Appealing to biologically plausible values, which are themselves subject to debate, does not resolve the problem, nor does the hand-tuning of thresholds based on visual inspection of PoG data, a procedure which assumes experimenter independence and expertise. A more objective method for classifying PoG data would therefore be desirable, especially if it were nonparametric and accommodated individual variation in a natural way.

Blignaut (2009) described an attempt to find an objective spatial threshold (the temporal threshold was maintained at 100 ms) for fixations extracted using dispersion-based techniques. A precise value was not successfully extracted, but a range of between 0.7 deg and 1.3 deg was judged to be acceptable, based on the consistency between implementations of their method using different dispersion metrics. Santella and DeCarlo (2004) outlined an interesting attempt to process eye movement data using a clustering technique based on the mean-shift procedure, which required the use of a kernel function. Typically a Gaussian kernel was used, and parameters controlling its scale were still selected by the user, meaning that the technique was not nonparametric. A method proposed by Engbert and Kliegle (2003) needed little user input, but it depended on a speed threshold being derived from the data by a method that, although data-driven, was not nonparametric in that it required a choice of multiplier of the estimated noise levels. The method was later extended by Nystrom and Holmqvist (2010), but retained its parametric dependence. Two other speed-based methods, one proposed by van der Lans et al. (2011) and one by Behrens et al. (2010) (which also used acceleration data) were also not nonparametric, requiring, respectively, a choice of minimum fixation duration and a choice of low-pass filter to smooth the velocity profile.

Clearly the existing methods for classifying fixations are unsatisfactory, and in Chapter 3 of this thesis an automatic and parameter-free method for classifying eye
movements which removes the problems of parameter selection is introduced.

1.5 Visual search

Being able to model search has great implications for the improvement of computer search algorithms (Elazary and Itti, 2010). Due to the inherent complexity of natural scenes, many of the early studies of visual search concentrated on discrete search arrays. Based on findings from experiments using search arrays, Treisman and Gelade (1980) developed the Feature-Integration Model of Attention. They described two types of visual search: feature search and conjunctive search. An example of a feature search is searching for a red square among green squares. In this situation, because the red square differs from distractors along one feature dimension (a feature dimension is defined by Treisman and Gormican (1988) as a set of mutually exclusive values for a single stimulus), it will ‘pop out’ and reaction time is independent of the number of distractors. Feature search is a consequence of parallel processing, which occurs in preattentive vision. An example of a conjunctive search is searching for a red circle among green circles and red squares. In this situation, the reaction time increases with the number of distractors, indicating that some manner of serial processing is taking place (Treisman and Gelade, 1980; Wolfe et al., 1989). In some situations it may be more difficult to define a search task as being either a feature search or a conjunction search. An example given by Treisman and Gelade (1980) was search for a blue circle among red circles and blue crosses. According to the Feature-Integration Model of Attention, this is a conjunction search, but if the display were divided into 15 red circles on the left and 15 blue crosses on the right an observer would be unlikely to scan serially through to the 30 items to find the target.
Treisman and Gelade’s (1980) Feature-Integration Model of Attention was based entirely on image properties. Wolfe (1994a) proposed an extension to the Feature-Integration Model of Attention called ‘Guided Search 2.0’. The updated model added a map quantifying the match of image properties to the target item to the map including the contribution of image features to the model. Wolfe (1994b) attempted to bridge the gap between artificial search arrays and natural scenes by generating artificial stimuli whose properties broadly resembled natural scenes and suggested that the Feature-Integration Model of Attention continued to apply with continuous stimuli.

Another approach to modelling visual search was taken by Koch and Ullman (1985), who presented a model of search with three stages. They suggested that in Stage 1, a set of elementary features is computed across the visual field to form a bottom-up saliency map, where a high value of salience means that a location was particularly interesting to the visual system. In Stage 2, a winner-take-all mechanism selects the location with the highest salience. In Stage 3, the properties of the selected location are routed to central representation. The model was not specifically designed to answer questions about visual search over natural scenes, but the idea of bottom-up saliency gave rise to Itti et al.’s (1998) biologically plausible saliency map, which is discussed in more detail in Section 1.6. Because humans performing visual search for an object in the real world make use of contextual information (for example, we would not search for car keys on the ceiling), Torralba et al. (2006) suggested that good model of visual search requires, in addition to image and target properties, the inclusion of information on scene ‘gist’. However, this is only true when the target has some contextual meaning.

More recent studies have attempted to establish what image properties in natural scenes affect the difficulty of a search task (Rosenholtz et al., 2007; Henderson et al., 2009). This issue will be discussed further in Chapter 4.
1.6 Predicting human eye movements

We increase the amount of information obtained when viewing a scene by moving our eyes, but there is a limit in how many locations we can fixate in a given time period, and when viewing a picture often there are areas which are never directly fixated. Because any gap in perception provides an opportunity for new compression algorithms, knowing more about where people look, and why, is of great interest.

Some of the earliest work in the area of eye movements when viewing complex scenes was undertaken by Yarbus (1967), who showed that the locations fixated by observers are strongly influenced by the task they are to undertake. Despite this, there have been various attempts to predict eye movements using visual saliency (introduced in the previous section). Parkhurst et al. (2002) showed that when observers were asked to freely view images of natural scenes, saliency levels calculated using Itti et al.’s (1998) model were higher at fixated locations than locations selected at random. Other saliency maps produced in an attempt to predict eye movements included a nonparametric saliency map, which was directly learned from eye movement data (Kienzle et al., 2006), a saliency map based on entropy (Kadir and Brady, 2001) and a saliency map based on the statistics of natural scenes (Zhang et al., 2008).

Although much attention has been paid in the literature to saliency maps, the extent to which they could be used to predict eye motion has perhaps been overestimated: as Einhäuser, Spain and Perona (2008) drily stated: “Some authors hope that, by progressively refining such low-level models, human attention will eventually be modelled perfectly”. Saliency is an image-based quantity, independent of the observer, and as such is denoted as a bottom-up factor. Torralba et al. (2006) highlighted the importance of top-down gaze-guidance factors. Top-down influences include factors such as scene gist and a viewer’s personal preferences. A simple example of a situation in which both of these factors might override purely bottom-up factors would be if an observer
were asked to view the image in Figure 1.1 and look for a duck floating on the water. It is intuitive that this might cause an observer to direct their attention to the water, where the duck is likely to be, rather than orienting attention towards the highly salient brightly coloured lanterns.

Opponents of the low-level saliency model of visual attention include Einhäuser, Spain and Perona (2008) and Cristino and Baddeley (2009). Einhäuser, Spain and Perona (2008) concluded that it is objects, not saliency, which predict the locations of fixations when observers view natural scenes, and that features thought to contribute to low-level saliency, such as edges, are simply correlates of objects. Cristino and Baddeley (2009) presented an eye-tracking study which made use of video stimuli. The videos were presented under different conditions, in which they had been temporally filtered. In this way it was possible to present stimuli which were identical in terms of high-level associations, but which differed in their low-level salience. They found that there was little difference between fixation distributions when the same video was viewed under different conditions, and concluded that image salience did not drive
fixations.

There is also debate about the timescale over which top-down and bottom-up factors operate. Some researchers have argued that top-down influences, such as task demands, can immediately override saliency effects (Einhäuser, Rutishauser and Koch, 2008). Other researchers have suggested that the influence of low-level features may be greatest at the earliest stages of image presentation (Itti and Koch (2001) suggests a timescale of 25-50 ms) becoming dominated by higher-level influences over time. Whether the influence of salience declines over time (Parkhurst et al., 2002) or if low-level features continue to contribute to an equal extent, with the influence of higher-level features increasing over time (Tatler et al., 2005), is a difficult question which has not yet been answered. However, there is evidence to support the hypothesis that bottom-up mechanisms do contribute to attentional guidance in at least some conditions (Parkhurst et al., 2002; Parkhurst and Niebur, 2004; Peters et al., 2005; Tatler et al., 2005).

1.7 Steering visual attention

Previous attempts to steer human visual attention using image modifications have achieved various levels of success. Approaches taken by previous studies have included both static image modifications such as increasing or decreasing contrast and temporal image modifications such as adding flicker. Some of these studies are summarized as follows.

1.7.1 Adjustment of local contrast

Reinagel and Zador (1999) showed that when freely viewing natural scenes, observers tended to fixate regions of high luminance contrast, defined as the standard deviation of the luminance values of pixels within a image region divided by the mean intensity
of the image. Following this, Einhäuser and König (2003) presented an experiment in which they attempted to modify local luminance contrast to investigate its effect on human scanpaths. It was found that contrast modification, be it an increase or reduction of local contrast, increased the probability that the modified region would be fixated. From this they concluded that high contrast itself does not have a causal influence upon scanpaths.

Although Einhäuser and König’s (2003) concept was sound, Parkhurst and Niebur (2004) highlighted several problems with their methodology, which could explain their results and conclusions. The major problem lay in the method used to modify local contrast: local contrast was adjusted by adding or subtracting a fraction of the difference between the greyscale value of a pixel and the average greyscale value across the entire image. If average local luminance were lower than average luminance of the whole scene, this attempt to reduce local contrast would actually have increased contrast. In addition to this, an adjustment of first-order luminance contrast on an image will cause a change in second-order texture contrast. Parkhurst and Niebur (2004) created a second-order saliency map, including texture contrast as a feature, and found that the inclusion of this feature would explain the results of Einhäuser and König (2003). A later paper (Baddeley and Tatler, 2006) described an attempt to establish which image features contribute to the allocation of visual attention suggested that it is edges, not simply ‘contrast’, which attract fixations. Açıkgöz et al. (2009) suggested that edges may guide attention where they are present, but in images which lack such properties, such as an image of a bush or undergrowth, lower level features such as contrast would influence attention.
1.7.2 Adjusting texture contrast

The findings of Parkhurst and Niebur (2004), motivated Su et al. (2005) to develop an image-processing technique for the adjustment of texture contrast. This technique was used to de-emphasize regions in natural scenes. Their method relied upon the manipulations of texture power maps, followed by a correction to remove any inadvertent adjustments made to first-order image properties. Validation of the success of the technique was performed by means of psychophysical experiments. In one experiment a search task was used to assess the effect of the modification. Observers were asked to search for a target object among a number of distractors. The search time for images where texture variance had been reduced everywhere except for the target was compared to the search time in unmodified images, and was indeed reduced for the modified image. An eye-tracking experiment was also performed in which observers were asked to freely view images for five seconds. A qualitative evaluation was made, but not a detailed analysis.

1.7.3 Guiding attention by geometry

Kim and Varshney (2008) investigated attentional guidance for use in the interpretation of volume visualizations involving the presentation of 3D data-sets, such as those obtained by magnetic-resonance scans. They based their definition of saliency on the geometry of the 3D mesh: mean curvature was calculated from the 3D mesh, and a centre-surround mechanism was used to identify regions which differed in curvature from their surroundings. These points were denoted ‘salient’. In order to guide attention, geometry was manipulated using smoothing and sharpening. The study found that the technique was able to attract a greater proportion of fixations to a region of interest than fell upon the same region in the original image.
1.7.4 Guiding attention by subtle flicker

Other methods that have been used to attract attention to selected areas in a natural scene include the application of a subtle luminance flicker effect to the desired region. McNamara et al. (2008) applied such a method to artificial stimuli and added their subtle flicker effect to regions containing small transparent balls. Observers were told to count the number of balls present in the stimuli under three conditions: no modulation of the stimuli, stimuli where the target locations had been modified by a subtle flicker effect which disappeared when observers directly viewed the stimuli and stimuli where flicker was made more obvious by using a larger modulation spread over a larger region. Observers only reported being aware of the image modification in the third condition. The performance of the observers differed between the groups significantly, with performance being improved when viewing the modified images.

Einhäuser, Rutishauser and Koch (2008) also attempted to increase saliency using flicker. Their flicker effect was achieved by increasing and decreasing contrast towards one side of an image at a rate of 5 Hz. They found that although observers would bias their first fixation after stimulus onset towards the flicker, it quickly ceased to attract attention. This does not necessarily undermine the use of flicker as a method of guiding attention; the flicker may have been too weak so it would not have a strong effect on attention, or it may have been too strong, so the image region would become unpleasant to look at.

1.7.5 Success in steering attention may depend on task

Einhäuser, Rutishauser and Koch (2008) reported a successful attempt to influence human scanpaths using properties thought to have an effect on image saliency. A contrast gradient was applied to increase saliency to either the left or the right hand side of the image. Observers were given one of two tasks to perform while viewing
the image. In one task, observers freely viewed images, and were asked whether reach image was ‘natural’, (i.e. representing a real-world scene or not); in another they were told to search for a ‘bullseye’ target hidden in the image. In the free viewing task, the fixations were biased towards the high-contrast side. When performing a search task, this bias disappeared immediately, suggesting that task demands had over-ridden low-level influences.

1.7.6 Image enhancement used in this project

Although there is some debate regarding the ability of an image modification to guide fixations, previous work suggests that gaze can be guided at least under some circumstances. Although contrast has in the past been linked to human attention, Baddeley and Tatler (2006) and Açıkl et al. (2009) have suggested that attention may be attracted by edges. In the experiments described in Chapter 5, images were modified using unsharp masking—a technique for increasing or decreasing the sharpness of edges—and the effect on human fixations was measured using an eye-tracker. In unsharp masking (Gonzalez and Wood, 1992), an image $I$ is processed using a low pass filter to produce a blurry version of the image $I'$, and the effect on human fixations was measured using an eye-tracker. In unsharp masking (Gonzalez and Wood, 1992), an image $I$ is processed using a low pass filter to produce a blurry version of the image $I'$. The blurred version is then subtracted from the original to produce a sharpened image $U$ as follows:

$$U = I + k(I - I'),$$  \hspace{1cm} (1.1)$$

where $k$ controls the level of enhancement. If $k$ were negative, the effect of the modification would be to reduce contrast, a process which will be referred to as de-enhancement. Figure 1.2 shows the effect of unsharp masking with $k > 0$.

In this project the contrast of images was adjusted locally. The constant $k$ in Equation 1.1 was replaced by a weight function $W$: 

Two different image modifications were used in this project. The first was a de-enhancement, which was weighted with a ramp function, where the level of de-enhancement smoothly decreased from the highest level at either the left or the right side of the image to no de-enhancement at the other side. The second was an enhancement with a Gaussian weighting function:

$$W(x, y) = we^{-[(x-x_c)^2+(y-y_c)^2]/2\sigma^2},$$  \hspace{1cm} (1.3)

where \((x_c, y_c)\) is the centre of the enhancement, \(w\) is its level and \(\sigma\) is the standard deviation of the enhancement.
1.8 Colour spaces

In Chapter 4 the effect of a number of image features upon the ability of human observers to detect a target are discussed. The images used in the experiment were in colour and the colour of surfaces varied across the scenes. In addition to image features such as luminance and edge density, colour was included in analysis. The measurement of colorimetric properties of a scene is a non-trivial task so a brief introduction to colour spaces is provided here.

**Colour mixing**

As has been mentioned earlier, colour vision in humans is achieved by comparing the outputs from the three different types of cones. Organisms which achieve colour perception in this way are known as trichromats, and the system itself is known as a trichromatic system. A consequence of the system is that many colours can be replicated by mixing the three primary colours, which permits, for example, a television with only three types of pixel to produce a realistic image of a real world scene containing a range of colours. The inability to distinguish between two stimuli despite their spectral differences is known as metamerism (Hunt, 1995). The level of each of the primary colours required to replicate a sample colour using additive mixing are known as the colour’s tristimulus values, and these values can be used to describe the colour of a particular light.

**CIE 1931**

Measuring the tristimulus values of a stimulus can be achieved using a bipartite colorimeter (Wyszecki and Stiles, 1967). The device is used to present an observer with a divided field of view. The stimulus is presented on one side of the field, and an additive mixture of the primary colours on the other. The observer adjusts the amount of
each primary colour until the mixture and the stimulus are identical, at which point the
tristimulus values can be read from the device. Using this method it is possible to de-
termine colour-matching functions $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$ that represent the level of each
of the three primary colours needed to match light of a given wavelength $\lambda$. To this
end, data were collected independently by two researchers: Wright (1929) and Guild
(1932). Occasionally some colours could not be reproduced by adjusting the intensity
of the primary lights being mixed. When this occurred one of the primary colours
could be transferred from the colour mixture to the stimulus under test and, once the
match had been made, the coefficient of this primary would be negative (Guild, 1932).
The two researchers also used different primaries, so their colour-matching functions
were different.

For computational convenience, the International Commission on Illumination (CIE)
transformed the colour-matching functions as produced by Wright (1929) and Guild
(1932) to a new set of standard non-negative functions $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$, producing
the CIE 1931 XYZ colour space (Wyszecki and Stiles, 1967). The values of $\bar{y}(\lambda)$ cor-
responds to the luminance sensitivity of the human eye. Using these co-ordinates, a
given colour $c(\lambda)$ can be represented by its tristimulus values $X$, $Y$, and $Z$, which are
obtained by integrating the product of $c(\lambda)$ with $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$, respectively.

**CIELAB**

Although CIE 1931 XYZ colour space can be used to reproduce colours using the
primary colours, it is perceptually non-uniform (Hunt, 1998). Two pairs of points can
be separated by the same Euclidean distance in XYZ space but, depending on their
locations, the perceptual difference can vary. Although not perceptually uniform, the
CIELAB colour space is more perceptually uniform than XYZ space (CIE, 2004).
CIELAB has co-ordinates $L^*, a^*,$ and $b^*$, which are produced from $X$, $Y$, and $Z$ co-ordinates (CIE, 2004) as follows:

$$L^* = \begin{cases} 
116(Y/Y_n)^{1/3} - 16, & \text{if } Y/Y_n > 0.008856, \\
903.3(Y/Y_n), & \text{if } Y/Y_n \leq 0.008856, 
\end{cases} \quad (1.4)$$

$$a^* = 500 \left[ f(X/X_n) - f(Y/Y_n) \right], \quad (1.5)$$

and

$$b^* = 200 \left[ f(Y/Y_n) - f(Z/Z_n) \right], \quad (1.6)$$

where

$$f(I) = \begin{cases} 
I^{1/3}, & \text{if } I > 0.008856 \\
7.787I + 16/116, & \text{if } I \leq 0.008856
\end{cases}$$

and $X_n$, $Y_n$ and $Z_n$ are the tristimulus values of a reference white. These values are usually obtained from a perfect reflecting diffuser illuminated by the same light source as the test scene (CIE, 2004). The co-ordinate $L^*$ represents lightness, $a^*$ represents a red-green axis and $b^*$ a yellow-blue axis (Westland and Ripamonti, 2004). The value of chroma is calculated (CIE, 2004) as follows:

$$C = \sqrt{(a^{*2} + b^{*2})}. \quad (1.7)$$

In this project images from a set of hyperspectral images were used. Hyperspectral images were used to ensure the accurate representation of scene colour when they were displayed on the computer monitor. The tristimulus values of the reference white were produced for the images by measuring the reflected spectrum from a small flat grey reference surface in the scene using a telespectroradiometer immediately after image acquisition (Foster et al., 2004).
1.9 Information theory

When attempting to steer human visual attention, it becomes necessary to compare gaze distributions. To provide a means of quantifying their similarity, information theory was investigated over the course of this project (see Chapter 2). The form in which information theory is most familiar was developed by Shannon (1948) to measure objectively the information carried in signals. Two terms associated with information theory will be introduced here: entropy, the measure of the disorder of a random variable (Cover and Thomas, 1991) which will be used in later derivations, and Kullback-Leibler (KL) divergence, the measure of the similarity of pairs of gaze distributions.

1.9.1 Entropy

If $X$ is a discrete random variable with probability mass function $p(x)$, its entropy $H$ is defined (Cover and Thomas, 1991) as

$$ H(X) = - \sum_{x \in X} p(x) \log p(x). \quad (1.8) $$

This quantity can be thought of as being the average amount of information required to describe this random variable.

The entropy of a continuous random variable can also be measured. If $X$ is a continuous random variable with probability density function $f(x)$, its differential entropy $h$ is defined (Cover and Thomas, 1991) as

$$ h(X) = - \int f(x) \log f(x) dx. \quad (1.9) $$

Differential entropy can be transformed to entropy by considering the result of
dividing the range of $X$ into bins of width $\Delta$ (Cover and Thomas, 1991). By the mean value theorem, there exists a value $x_i$ within each bin such that

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx. \tag{1.10}$$

Consider the quantized random variable $X^\Delta$, which is defined as

$$X^\Delta = x_i, \text{ if } i\Delta \leq X < (i + 1)\Delta. \tag{1.11}$$

The probability that $X^\Delta = x_i$ is

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(i\Delta). \tag{1.12}$$

The entropy of $X^\Delta$ can then be expressed as follows

$$H(X^\Delta) = -\sum_{-\infty}^{\infty} p_i \log p_i,$$

$$= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta),$$

$$= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)) - \sum_{-\infty}^{\infty} f(x_i)\Delta \log(\Delta),$$

$$= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)) - \log \Delta, \tag{1.13}$$

As $\Delta \to 0$, the first term in Equation 1.13 tends towards $\int f(x) \log f(x), dx$, so

$$H(X^\Delta) + \log \Delta \to h(f) = h(X) \text{ as } \Delta \to 0. \tag{1.14}$$
CHAPTER 1. INTRODUCTION

1.9.2 KL divergence

KL divergence, also known as ‘relative entropy’, is an information-theoretic measure of the difference between two probability distributions. It is defined for probability mass functions \( p \) and \( q \) (Cover and Thomas, 1991) as

\[
d_{KL}(p, q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}
\]

(1.15)

and for probability density functions \( r \) and \( s \) (Cover and Thomas, 1991) as

\[
d_{KL}(r, s) = \int r(x) \log \frac{r(x)}{s(x)} dx.
\]

(1.16)

The KL divergence is zero only if the two probability distributions are identical, otherwise \( d_{KL}(p, q) > 0 \). One undesirable aspect of KL divergence is that it is an asymmetric quantity: \( d_{KL}(p, q) \) does not necessarily equal \( d_{KL}(q, p) \). This problem can be solved by reporting the symmetric quantity \( d_{KL}(p, q) + d_{KL}(q, p) \) (this is actually the form in which the quantity was introduced by Kullback and Leibler (1951)). The accurate measurement of KL divergence is a non-trivial task, and is discussed in more detail in Chapter 2.

1.10 Thesis outline

The main objectives of this project have been introduced and the background information summarized. The order in which they are examined in more detail is as follows:

**Chapter 2: Kullback-Leibler divergence.** There are a number of difficulties in the measurement of KL divergence. In Chapter 2 these problems are outlined and a number of estimators tested and compared in order to identify the best KL estimator for
comparing fixation distributions. It was found that a method for estimating Kullback-Leibler Divergence based upon nearest-neighbours outperformed existing methods.

Chapter 3: Collection and analysis of eye movement data. Data were collected over the course of this project using a Cambridge Research Systems (CRS) High-Speed Video Eyetracker Toolbox. The device and calibration routines were tested using both human observers and an artificial eye. In response to the problems in fixation classification described in Section 1.4 a novel parameter-free method (Mould et al., 2012) for classifying eye movements is proposed. The proposed method was verified against scanpaths which had been classified by experts.

Chapter 4: Searching images of natural scenes. An experiment was undertaken during which observers were required to search for a target superimposed on natural scenes in order to establish why they might detect the target in some trials and not others. It was found that of the image factors investigated, which included chroma, luminance, edge density and various measures of image ‘clutter’, chroma was most important to the detection of the target.

Chapter 5: Steering human fixations. The experiments described in Chapter 5 aimed to establish whether the fixations of observers could be guided using image modifications. Two different image modifications were tested, both based on unsharp masking. It was found that an image modification which reduced edge sharpness could not be used to steer fixations away from large image regions, but that fixations could be attracted towards an image region which had been locally enhanced. The local enhancement was also shown to have improved detection ability when applied near to a target.
**Chapter 6: Conclusions and further work.** A summary of the salient points in the thesis is presented, with suggestions for further work.
Chapter 2

Kullback-Leibler Divergence

In this chapter, the estimation of KL divergence (which was introduced in Section 1.9.2) will be discussed. KL divergence is an information theoretic quantity that has previously been used to compare fixation distributions (Frey et al., 2008). However, calculating any information-theoretic estimate is not a straightforward task. For example, if data are binned in order to calculate entropy (as defined in Section 1.9.1), statistical fluctuations make the distribution less uniform, and entropy is underestimated (Grassberger, 2008). This problem has been solved by estimating entropy directly from the data rather than estimating probability distributions before estimating entropy using a nearest-neighbours approach (Kozachenko and Leonenko, 1987). The estimator described by Kozachenko and Leonenko (1987) was implemented by Iván Marín-Franch and its performance is shown in Figure 2.1.

As with entropy there are difficulties when attempting to estimate KL divergence by binning data, particularly when the data are undersampled. Insufficient samples can lead to empty bins and due to the form of Equation 1.15, it is clear that an empty bin in either histogram could cause numerical problems. The conventions $0 \log \frac{0}{q} = 0$ (which is justified by continuity) and $p \log \frac{p}{0} = \infty$ are generally used when estimating KL divergence (Cover and Thomas, 1991). The former convention presents a problem, as
it violates the assertion that KL divergence should only be zero if calculated between two identical distributions; the latter convention presents a numerical problem. The problem of undersampling can only avoided when $N >> M$ ($N$ being the sample size and $M$ being the number of bins), a condition which is rarely satisfied. Attempts will be made to solve these problems in this chapter, and three methods for estimating KL divergence will be discussed in the following.

2.1 Methods of estimating KL divergence

One method of measuring the KL divergence between two data sets is to first estimate the probability mass function by binning the data into histograms and input the result to Equation 1.15. Another is to estimate the probability mass function using a kernel density estimator, with a kernel such as a Gaussian function, and again input the result
to Equation 1.15. Either of these methods can give points with zero density, which, as has been highlighted, can cause numerical problems. Some researchers (Tatler et al., 2005, for example) avoid the situation by adding a small prior to all bins before normalizing, but its impact is unclear. The effect of the size of the prior will be tested in Sections 2.1.2 and 2.1.3. When data are sampled from continuous distributions, it is also necessary for the user to define the resolution of the probability mass functions. The dependence of estimates upon this quantity is also tested in Sections 2.1.2 and 2.1.3.

A different type of method would be to measure the KL divergence directly from data, without the intermediate step of estimating probability mass functions. Such a method was provided by Leonenko et al. (2008), and is tested in Section 2.2. An attempt to improve its performance will also be described in the same section.

2.1.1 Synthetic data set

The three methods for measuring KL divergence mentioned above were tested using synthetic data sets sampled from two different 2-dimensional distributions composed of arbitrary superpositions of Gaussian functions. Testing was performed using 2-dimensional data because gaze data are typically 2-dimensional distributions, but the principles could be extended to $N$-dimensional data. A Gaussian distribution of mean $\mu$ and standard deviation $\sigma$ is written here as $N(\mu, \sigma)$. One data set was sampled from a superposition of $N([0, 0], [2, 2]), N([3, -3], [2, 2])$, and $N([-3, 2], [2, 2])$, and the other from a superposition of $N([-3, -3], [2.5, 2.5]), N([-2.5, -3], [2.5, 2.5])$, and $N([2.5, 1], [2.5, 2.5])$. The first two methods required the data to be sampled from a finite range, so the data were clipped in each dimension to a range of -5 to 5. No analytical solution was derived for the KL divergence between these two distributions (although an analytical solution could be derived for the KL divergence between
Figure 2.2: KL divergence between two distributions composed of superpositions of Gaussian functions (see Section 2.1.1 for details) calculated from a discrete approximation to the underlying probability density function as the number of bins was varied.

untrimmed distributions, see Section 2.2), but it was possible to estimate the KL divergence between the underlying distributions accurately using a numerical method. This was achieved by dividing the range over which the distributions existed into a number of discrete bins $B$ and calculating the probability mass function for each bin, to form a discrete approximation of the probability density function. This probability mass function was then input to Equation 1.15. As $B \to \infty$, the estimated value quickly converged towards 0.9763 bits (results are shown in Figure 2.2). Although the KL divergence estimate rapidly converged, notice that even when estimating KL divergence from underlying probability distributions rather than sampled data, there is (small) bias evident when the resolution is varied. This shows that even when removing any confounds associated with sample size, using binning methods to estimate KL divergence can result in biased estimates.
Figure 2.3: KL divergence between two data sets sampled from distributions composed of superpositions of Gaussian functions (see Section 2.1.1 for details) estimated using the method described in Section 2.1.2 as $N$ and $p$ were varied and $r$ was set to 20. Results shown are the means taken over 100 runs.

2.1.2 Binning and adding a constant

The method of KL divergence estimation by binning data and adding a constant has two user-defined parameters, namely the size of the prior $p$ and the resolution of the histogram $r$. The effect of these values upon the estimated KL value were tested using data sampled from the arbitrary distributions detailed in Section 2.1.1 while the size of the data set was varied. Data were binned into a histogram of resolution $r \times r$ and a prior $p$ was added to each bin. The histograms were then normalized, and Equation 1.15 used to calculate the KL divergence.

First the resolution of the histogram was fixed at $r = 20$ and the prior was varied over $p = 10^{-15}, 10^{-12}, ..., 10^0$. The sample size was varied over $N = 10^2, 10^{2.5}, 10^3, ..., 10^6$. Figure 2.3 shows that varying $p$ can have a great effect on the KL divergence, but that this effect is reduced when $N$ is increased.
Next the size of the prior was fixed to $p = 10^{-6}$ and the resolution varied over $r = 5, 10, 20, ..., 80$. The estimated KL divergence values are shown in Figure 2.4, and reveal that if the number of bins is too small, the estimates can converge to an incorrect value. However, if the resolution is too high, the rate of convergence is reduced.
Figure 2.4: KL divergence between two data sets sampled from distributions composed of superpositions of Gaussian functions (see Section 2.1.1 for details) estimated using the method described in Section 2.1.2 as $N$ and $r$ were varied and $p$ was set to $10^{-6}$. The upper graph shows the full range of the data, and lower graph has a narrower vertical scale to show detail. Results shown are the means taken over 100 runs.
2.1.3 Kernel smoothing

Probability mass functions can also be estimated by a kernel density estimator. Using the data sampled from the same distributions as the previous section, a state-of-the-art 2-dimensional kernel density estimator which optimally smooths data over a grid of a user-specified resolution with a bivariate Gaussian kernel (Botev et al., 2010) was used to estimate the probability mass functions and, as in Section 2.4, a small prior was added to the resulting histograms. The histograms were then normalized, and Equation 1.15 used to calculate the KL divergence.

First the resolution of the histogram was fixed at \( r = 2^8 \) and the prior was varied over \( p = 10^{-15}, 10^{-12}, ..., 10^0 \). The sample size was varied over \( N = 10^2, 10^2.5, 10^3, ..., 10^6 \). Figure 2.5 shows that for \( p \leq 10^{-6} \), the KL estimates show considerably more stability than those shown in Figure 2.3.

Next the size of the prior was fixed to \( p = 10^{-6} \) and the resolution varied over \( r = 2^5, 2^6, ..., 2^{11} \). The estimated KL divergence values are shown in Figure 2.6, and suggest that the resolution of the grid has less effect on the calculated values than when data are crudely binned as in Section 2.1.2.
Figure 2.5: KL divergence between two data sets sampled from distributions composed of superpositions of Gaussian functions (see Section 2.1.1 for details) estimated using the method described in Section 2.1.3 as $N$ and $p$ were varied and $r$ was fixed to $2^8$. Results shown are the means taken over 100 runs.
Figure 2.6: KL divergence between two data sets sampled from distributions composed of superpositions of Gaussian functions (see Section 2.1.1 for details) estimated using the method described in Section 2.1.3 as $N$ and $r$ were varied and $p$ was set to $10^{-6}$. Results shown are the means taken over 100 runs.
2.1.4 Leonenko et al.’s KL divergence estimator

An alternative to estimating the underlying probability distribution from data which may be sparsely sampled, as in Sections 2.1.2 and 2.1.3, before estimating KL divergence would be to estimate KL divergence directly from the data itself. Leonenko et al. (2008) outlined an estimator for KL divergence based upon nearest-neighbour statistics, which was implemented by Iván Marín-Franch. Since the algorithm takes as input only the data points, not an approximation to the underlying distribution, parameters such as histogram resolution or size of prior are not required. This algorithm, and an offset version of the algorithm (explained in Section 2.2.3) were used to estimate KL divergence between data sets sampled from the same distributions as those in the previous two sections. Results are shown in Figure 2.7, and show that Leonenko et al.’s (2008) estimator converges quickly to the true KL divergence value. In addition to this, the lack of any user-defined parameters means that the instability seen in Figures 2.3, 2.4, 2.5 and 2.6 does not present a similar problem here.
Figure 2.7: KL divergence between two distributions composed of superpositions of Gaussian functions (see Section 2.1.1 for details) estimated using Leonenko et al.’s (2008) estimator as $N$ was varied. Results shown are the means taken over 100 runs.

2.2 Testing of Leonenko et al.’s KL divergence estimator against an analytical solution

Since Leonenko et al.’s (2008) algorithm estimates KL divergence from the data points and no binning process is required, the data-range need not be specified. This makes it possible to input data sampled from known distributions such as Gaussian functions, and compare the KL divergence as calculated by the algorithm with that calculated using an analytical expression.
2.2.1 Analytical expression for KL divergence between two Gaussian distributions

No analytical expression for the KL divergence between two Gaussian functions of non-zero means could be found in the literature so an expression was derived as follows.

Let $\phi_p(x)$ and $\phi_q(x)$ represent two $D$-dimensional Gaussian distributions with variance-covariance matrices $P$ and $Q$, and means $\mu_p$ and $\mu_q$ respectively. KL divergence can be expressed as follows

$$d_{KL}(\phi_p||\phi_q) = \int \phi_p \ln \frac{\phi_p}{\phi_q} \, dx$$

$$= \int \phi_p \ln \phi_p \, dx - \int \phi_p \ln \phi_q \, dx$$

$$= \bar{h}(\phi_p, \phi_q) - h(\phi_p), \quad (2.1)$$

where $\bar{h}$ is termed cross entropy (Lawrence, 2005) and $h$ is differential entropy.

Although an analytical expression for $h$ was available (Ahmed and Gokhale, 1989), the following derivation is given as preparation for the derivation of an expression for $\bar{h}$:
\[ h = - \int \phi_p \ln \phi_p \, dx \]
\[ = - \int \phi_p \ln \left( \frac{1}{(2\pi)^{D/2} |\mathbf{P}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x} - \mu_p)'\mathbf{P}^{-1}(\mathbf{x} - \mu_p)} \right) \, dx \]
\[ = \int \phi_p \, dx \cdot \ln \left( (2\pi)^{D/2} |\mathbf{P}|^{1/2} \right) - \int \phi_p \ln e^{-\frac{1}{2}(\mathbf{x} - \mu_p)'\mathbf{P}^{-1}(\mathbf{x} - \mu_p)} \, dx \]
\[ = \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |\mathbf{P}| + \frac{1}{2} \int \phi_p (x - \mu_p)'\mathbf{P}^{-1}(x - \mu_p) \, dx \]
\[ = \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |\mathbf{P}| + \frac{1}{2} \mathbb{E}[(X - \mu_p)'\mathbf{P}^{-1}(X - \mu_p)], \]

where \( X \sim \mathcal{N}(\mu_p, \mathbf{P}) \).

For a random vector \( Y \sim \mathcal{N}(\mu, \sigma) \) and positive definite matrix \( A \) (Sahai and Ojeda, 2005),

\[
E(Y'AY) = \text{tr}(A\sigma) + \mu' A \mu. \tag{2.2}
\]

And since \( (X - \mu_p) \sim \mathcal{N}(0, \mathbf{P}) \):

\[
\mathbb{E}[(X - \mu_p)'\mathbf{P}^{-1}(X - \mu_p)] = \text{tr}(\mathbf{P}^{-1}\mathbf{P}) = D.
\]

Therefore, the analytical expression of \( h \) is

\[ h = \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |\mathbf{P}| + \frac{D}{2}. \]

The result agrees with the equation for entropy of a Gaussian function given by Ahmed and Gokhale (1989).

An analytical expression for \( \bar{h} \) was also derived:
\[ \hat{h} = - \int \phi_p \ln \phi_q \, dx \]
\[ = - \int \phi_p \ln \frac{1}{(2\pi)^{\frac{D}{2}} |Q|^{\frac{1}{2}}} e^{-\frac{1}{2}((x-\mu_q)Q^{-1}(x-\mu_q))} \, dx \]
\[ = \int \phi_p \, dx \cdot \ln(2\pi)^{\frac{D}{2}} |Q|^{\frac{1}{2}} - \int \phi_p \ln e^{-\frac{1}{2}((x-\mu_q)Q^{-1}(x-\mu_q))} \, dx \]
\[ = \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |Q| + \frac{1}{2} \int \phi_p (x-\mu_q)^T Q^{-1} (x-\mu_q) \, dx \]
\[ = \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |Q| + \frac{1}{2} E[(X-\mu_q)^T Q^{-1} (X-\mu_q)], \]

where \( X \sim \mathcal{N}(\mu_p, P) \)

Because \( (X-\mu_q) \sim \mathcal{N}(\mu_p - \mu_q, P) \), Equation 2.2 can again be used:

\[ E[(X-\mu_q)^T Q^{-1} (X-\mu_q)] = \text{tr}(Q^{-1}P) + (\mu_p - \mu_q) Q^{-1} (\mu_p - \mu_q), \]

and

\[ \hat{h} = \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |Q| + \frac{1}{2} \text{tr}(Q^{-1}P) + \frac{1}{2} (\mu_p - \mu_q) Q^{-1} (\mu_p - \mu_q), \]

and finally

\[ d_{KL}(\phi_p || \phi_q) = \hat{h} - h = \frac{1}{2} \left[ D \ln 2\pi + \ln |Q| + \text{tr}(Q^{-1}P) + (\mu_p - \mu_q) Q^{-1} (\mu_p - \mu_q) \right] \]
\[ - \left( \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |P| + \frac{D}{2} \right). \]

This reduces to:

\[ d_{KL}(\phi_p || \phi_q) = \frac{1}{2} \left[ \ln \frac{|Q|}{|P|} + \text{tr}(Q^{-1}P) + (\mu_p - \mu_q) Q^{-1} (\mu_p - \mu_q) - D \right]. \quad (2.3) \]
The KL divergence between Gaussian distributions with zero (or equal) means is given by Lawrence (2005), and is shown here as Equation 2.4.

\[
d_{KL}(\phi_p || \phi_q) = \frac{1}{2} \left[ \ln \frac{|Q|}{|P|} + \text{tr}(PQ^{-1}) - D \right].
\] (2.4)

Clearly if \(\mu_p = \mu_q\), Equation 2.3 is equivalent to Equation 2.4.

**2.2.2 Use of Leonenko et al. estimator to calculate KL divergence between two Gaussian functions**

Leonenko et al.’s (2008) estimator was used to calculate the KL divergence between data sampled from a pair of 2-dimensional Gaussian functions \(\mathcal{N}([0, 0], [2, 2])\) and \(\mathcal{N}([0, 2], [3, 3])\). The sample size was identical for both, and varied over \(N = 10^2, 10^2.5 \ldots, 10^6\). The results are shown in Figure 2.8 and show the estimates to converge to the true KL value.

**2.2.3 The offset version of the Leonenko et al. estimator**

Marín-Franch (2009) showed that it is possible to improve an estimator of mutual information by partitioning the values into their Gaussian and non-Gaussian components. Similar attempts were made here to find an offset estimator for the KL divergence.

As mentioned earlier, KL divergence can be written as \(d_{KL}(R, S) = \bar{h}(R, S) - h(R)\), where \(h\) is entropy and \(\bar{h}\) is termed cross entropy (Lawrence, 2005). A property of entropy is that if a continuous random variable \(R\) is subject to an invertible linear transform \(t\), i.e. \(R' = tR\), the differential entropy of \(R\) and \(R'\) are related as follows (Cover and Thomas, 1991)

\[
h(R) = h(R') - \log |t|,
\] (2.5)

where \(|t|\) represents the absolute value of the determinant of \(t\).
Figure 2.8: Kullback-Leibler divergence between two data sets sampled from distributions $N([0, 0], [2, 2])$ and $N([0, 2], [3, 3])$ estimated using Leonenko et al.’s (2008) estimator, and the offset version of the same estimator (explained in Section 2.2.3) as $N$ was varied. The values shown are the means of 100 simulations.
If \( R' = (\text{Var}(R))^{-1/2} R \), then (Marín-Franch, 2009)

\[
h(R) = h(R') + \frac{1}{2} \log |\text{Var}(R)|. \tag{2.6}
\]

The term \( \frac{1}{2} \log |\text{Var}(R)| \) is equal to the differential entropy of a Gaussian with the same variance as \( R \) (Cover and Thomas, 1991) and so the Gaussian component of the entropy of \( R \) has been separated from the non-Gaussian component, \( h(R') \).

An analogous expression to Equation 2.5, which is concerned with a transform applied to cross entropy rather than differential entropy, was also required, and derived as follows.

Suppose that the two random variables \( R \) and \( S \) have pdfs \( f_p(x) \) and \( f_q(x) \) and are subject to an invertible linear transform \( t \), so \( R' = tR \) and \( S' = tS \), where \( R' \) and \( S' \) exist over a space \( y \). According to change of variable theorem (Poirier, 1995), the pdfs are then transformed (Cover and Thomas, 1991) as follows

\[
f_{r'}(y) = \frac{1}{|t|} f_r\left(\frac{y}{t}\right) \tag{2.7}
\]

and

\[
f_{s'}(y) = \frac{1}{|t|} f_s\left(\frac{y}{t}\right). \tag{2.8}
\]

The cross entropy of the two transformed random variables can then be expressed as

\[
h(R', S') = - \int \frac{1}{|t|} f_r\left(\frac{y}{t}\right) \log \frac{1}{|t|} f_s\left(\frac{y}{t}\right) dy.
\]

Since \( dy/dx = t \), we can reduce this equation as follows
\[
\hat{h}(R', S') = -\int \frac{1}{|\eta|} f_r \left( \frac{\eta}{\sqrt{\gamma}} \right) \log \left( \frac{1}{|\eta|} f_s \left( \frac{\eta}{\sqrt{\gamma}} \right) \right) \, d\eta
\]
\[
= -\int f_r(x) \log \left( \frac{1}{|\eta|} f_s(x) \right) \, dx
\]
\[
= -\int f_r(x) \log f_s(x) \, dx + \int f_r(x) \log |\eta| \, dx
\]
\[
= \hat{h}(R, S) + \log |\xi|.
\]

Analogous to the transform of differential entropy (Equation 2.5), we have a transform of differential cross entropy:

\[
\hat{h}(R, S) = \hat{h}(R', S') - \log |\xi|.
\] (2.9)

The offset version of the estimator of cross entropy would be written as

\[
\hat{h}(R, S) = \hat{h}(R', S') + \hat{h}_{\text{EG}}(R, S),
\] (2.10)

where \(\hat{h}_{\text{EG}}(R, S)\) is the cross entropy of two Gaussian functions with the same variance matrices as \(R\) and \(S\), i.e. the cross entropy of their equivalent Gaussian distributions.

As discussed earlier (Section 2.2.1), the expression for the cross entropy between two zero-mean \(D\)-dimensional Gaussian distributions with variance matrices \(R\) and \(S\) is

\[
\hat{h}(\mathcal{N}(0, R), \mathcal{N}(0, S)) = \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |S| + \frac{1}{2} \text{tr} \left( S^{-1} R \right).
\] (2.11)

Substituting Equation 2.11 into Equation 2.10 gives
\[
\bar{h}(R, S) = \bar{h}(R', S') + \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |S| + \frac{1}{2} \text{tr} \left( S^{-1} R \right)
\]
\[
= \bar{h}(R', S') + \ln(2\pi)^{D/2} + \ln |S|^\frac{1}{2} + \ln \left[ \exp \left( \frac{1}{2} \text{tr} \left( S^{-1} R \right) \right) \right]
\]
\[
= \bar{h}(R', S') + \ln \left[ (2\pi)^{D/2} |S|^\frac{1}{2} \exp \left( \frac{1}{2} \text{tr} \left( S^{-1} R \right) \right) \right]
\]
\[
= \bar{h}(R', S') - \ln \left[ (2\pi)^{D/2} |S|^\frac{1}{2} \exp \left( \frac{1}{2} \text{tr} \left( S^{-1} R \right) \right) \right]^{-1},
\]
so the transformation is \( t = \left( (2\pi)^{D/2} \exp \left[ \frac{1}{2} \text{tr} \left( S^{-1} R \right) \right] S^{\frac{1}{2}} \right)^{-1} \), or \( S^{-\frac{1}{2}} \) multiplied by a constant.

The offset estimator was used when estimating KL divergence for the data in Figures 2.7 and 2.8, and revealed very little improvement in the estimator. This is possibly because the transform is dependent upon only one of the two input distributions, unlike the offset version of the estimator of mutual information introduced by Marín-Franch (2009).

### 2.3 Behaviour of Leonenko et al.’s (2008) KL estimator with data sampled from identical distributions

Leonenko et al.’s (2008) KL estimator was also tested on data sampled from two identical distributions. As has already been stated, this is the only condition under which the KL divergence is zero and Figure 2.9 shows that the estimator once again converges to the correct value. As evidenced in Figure 2.9, the estimator does occasionally give small negative KL values, which, as was explained earlier, cannot be correct. This occurrence is not indicative of a fundamental problem with the estimator; it is simply a consequence of the estimated entropy of the distributions exceeding the estimated cross
entropy between them (see Equation 2.1) because the former is being overestimated or the latter is being underestimated, or both.

2.4 Discussion

KL Divergence provides a nonparametric method for comparing two sets of fixation distributions, but the most accurate method for estimating its value was unclear. To address this concern, three methods of estimating KL Divergence have been tested in this chapter:

1. Estimating the probability density function by binning data, adding a prior and using the discrete form of the KL divergence equation.

2. Estimating the probability density function using kernel density estimation, adding
a prior and using the discrete form of the KL divergence equation.

3. Estimating the KL divergence directly from the data using the algorithm developed by Leonenko et al. (2008).

Methods 1 and 2 differ from Method 3 in that they require user-defined input parameters. The input parameters were shown in Figures 2.3, 2.4, 2.5 and 2.6 to have an unpredictable effect on estimates including different rates of convergence and convergence to an inaccurate value. Method 3 is nonparametric so was not subject to similar problems, and was shown in Figure 2.7 to converge to the true KL value more quickly than Methods 1 and 2. Further testing of Method 3 using a different data set (data shown in Figure 2.8) showed that it converged correctly to an analytically calculated KL value. Attempts were made to improve the estimator by separating the Gaussian and non-Gaussian components of the estimate in a manner that had previously been shown to improve estimates of mutual information (Marín-Franch, 2009). This offset version of the estimator was shown in Figures 2.7, 2.8 and 2.9. The modification showed a little improvement to the estimator in Figure 2.7.

Because of its accuracy, rate of convergence and lack of sensitivity to user-defined parameters, the original form of Leonenko et al.’s (2008) estimator is used to estimate KL divergence in this thesis.
Chapter 3

Collection and Analysis of Eye Movement Data

The eye-tracker used in this project was tested using both an artificial eye and human observers and details of the tests are given in Section 3.1. In Section 1.4 a number of problems in the objective classification of fixations were outlined. A method designed to solve these problems is described in Section 3.2. The accuracy of the method was tested by comparisons with fixations labelled by experts. The accuracy of two existing parametric methods was tested in the same way for comparison purposes. The robustness of the proposed method to instrumental noise was verified in a separate simulation.

3.1 The CRS Video Eyetacker Toolbox

The eye-tracker used in this project was a monocular video eye-tracker running at 250 frames per second (High Speed Video Eyetracker Toolbox Mk 2; Cambridge Research Systems Ltd, UK) which recorded point-of-gaze (PoG) by monitoring the positions of two glints of light on the corneal surface, and the position of the pupil, as shown
in Figure 3.1. The relative position of these reference points on the eye provided information on the direction of gaze of the eye. In order to accurately convert this quantity into a PoG relative to the monitor, it was necessary for observers to undertake a calibration procedure each time the eye-tracker was used. If either of the corneal glints or pupil became obscured (for example by an eyelid) or if the image of the eye moved out of focus because of observer movement, tracking could be lost. Other possible sources of error in the eye tracker include the fact that the centre of the pupil may not remain fixed relative to the eyeball as the pupil dilates or constricts (Wyatt, 2010) or blurring of the two corneal glints.

### 3.1.1 Testing with Artificial Eye

An objective measurement of the eye-tracker’s accuracy was made using an artificial eye. It was necessary to test the eye-tracker using an artificial eye which could be pointed at known orientations. The use of an artificial eye also removed any confounding factors due to involuntary movements of a living eye (see Section 1.4).

**Apparatus**

A white wooden ‘eye’ with a hollow pupil was constructed. The inside of the hollow pupil was painted with a black acrylic paint with a non-glossy finish and a hard clinical contact lens fitted to the front of the eye to act as the cornea. The eye-tracker was able...
CHAPTER 3. COLLECTION AND ANALYSIS OF EYE MOVEMENT DATA

Figure 3.2: A screen shot of the artificial eye being tracked by monitoring the centre of the pupil and the corneal glints.

Figure 3.3: The Vernier scale used to adjust the artificial eye.

to track the artificial eye as it would a human eye (Figure 3.2). A mount for the artificial eye was constructed by the workshop in the School of Electrical and Electronic Engineering (University of Manchester) which allowed the eye to be rotated vertically and horizontally by means of a Vernier scale with a resolution of 5 arcminutes. The mount is shown in Figure 3.3.

Procedure

The artificial eye was moved to each of twenty orientations, arranged in a $5 \times 4$ grid with points evenly spaced between $-9$ deg and 9 deg horizontally and $-6$ deg and 9 deg vertically, and the PoG measured by the eye-tracker for each point. The measured horizontal and vertical gaze positions $x_m$ and $y_m$ were mapped to the corresponding orientations $x_t$ and $y_t$ by linear transformations,

$$x_t = ax_m + by_m + c$$ (3.1)
and

\[ y_t = dx_m + ey_m + f, \]  

(3.2)

where the constants \(a, b, c, d, e\) and \(f\) were optimized to give least-squares error. These transformations could then be used to map all measured PoG signals to screen co-ordinates.

After calibration data had been recorded, PoG of the artificial eye was measured as the eye was moved over horizontal angles \(-9\) deg, \(-8\) deg, ..., \(9\) deg, with the vertical angle fixed at 0 deg, then moved over vertical angles \(-6\) deg, \(-5\) deg, ..., \(9\) deg, with the horizontal angle fixed at 0 deg (outside of these ranges the glints had moved off the pupil and tracking failed).

**Results**

The RMS error of the calibration for the artificial eye was \(0.23 \pm 0.02\) deg (mean \(\pm\) SE) and the RMS error as the eye was pointed at the various orientations was \(0.18 \pm 0.01\) deg. These values give an indication of the precision of the eye-tracker itself (without any errors attributable to the observers).

### 3.1.2 Calibration and testing with human observers

The artificial eye was used to produce an objective error measure of the eye-tracker, but since the eye-tracker was used in practice to measure the eye movements of human observers, the eye-tracker was also tested using human observers. Data were recorded while observers fixated targets of known positions. It was possible that an observer’s head may drift over the course of a block. To test if such an effect could be accounted for calibration data were recorded both before and after the trials, and the transformation used to map the data obtained during the trials to screen coordinates was obtained in three different ways, two of which would account for an observer’s drift (if any).
CHAPTER 3. COLLECTION AND ANALYSIS OF EYE MOVEMENT DATA 71

Apparatus

The eye-tracker was described at the beginning of this section. Images were displayed on a 20-inch RGB CRT colour monitor (GDM-F520, Sony Corp., Tokyo, Japan) controlled by a graphics workstation (Fuel V12, Silicon Graphics Inc. Mountain View, CA). The spatial resolution of the display was 1600 × 1200 pixels; the intensity resolution on each RGB gun was 10 bits; and the refresh rate was approx. 60 Hz. The display subtended approximately 17 deg × 13 deg at a viewing distance of 1 m.

Observers

Three observers took part in the experiment (aged 22 to 29 years, two male, one female). All had normal colour vision and normal or corrected-to-normal visual acuity.

Procedure

For the calibration routine, observers fixated 20 calibration targets (0.26 × 0.26 deg black crosses), which were arranged in a 5 × 4 grid and presented sequentially on the screen at known positions. Observers were asked to click a mouse button as they fixated each target, and the corresponding horizontal and vertical PoG signals were recorded by the eye-tracker. After the calibration, the transformations between the twenty recorded PoG signals and corresponding calibration target positions were optimized by minimizing squared error. If the least-squares error exceeded 0.30 deg, the calibration was repeated. Calibration data were again recorded at the end of each block, producing two complete sequences of twenty measured gaze positions.

Each experimental block contained forty trials. In each trial observers were presented with a target (identical to the calibration target). Targets appeared twice at each of twenty locations, in a random order. Observers were required to fixate the target,
and click the mouse button when they were fixating the target steadily (as in the calibration). The mouse click prompted the onset of next trial. In each block there were forty trials and each observer undertook two blocks.

The three methods for obtaining the transformation used to map the data obtained during the experimental trials to screen co-ordinates were as follows:

1. The transformation was optimized using the first set of calibration data alone.
2. The mean was taken of the two sets of calibration data pointwise, and the resulting set of twenty calibration points used to optimize the transformation.
3. A weighted mean of the two sets of calibration data was calculated for each of the forty trials. The transformations were optimized for each trial individually.

Methods 2 and 3 were intended to account for an observer’s drift (if any).

**Results**

Method 1 gave an RMS error on gaze locations over all trials of $0.46 \pm 0.01$ deg.
Method 2 gave $0.49 \pm 0.02$ deg and Method 3 gave $0.51 \pm 0.02$ deg. These numbers give an indication of the level of error that can be expected from eye movement data recorded by this eye-tracker. The results also suggest that Methods 2 and 3 gave no great advantage in accounting for drift. The calibration routines were improved slightly over the course of this project, so will be explained for each experiment individually.

### 3.2 Classification of fixations

As has already been mentioned in Section 1.4, when viewing static scenes eye movements can be broadly classified as being either saccadic or fixational and there exists no parameter-free method for achieving this classification. This is a problem as it has
been reported that the method used to classify eye movement data can have a great impact on the reported fixations (Shic et al., 2008). The findings of Shic et al. (2008) were replicated by testing the effect of varying the thresholds used by two existing parametric methods for classifying fixations while counting the number of extracted fixations (Sections 3.2.2 and 3.2.3). One method was a stability-based method due to van der Linde et al. (2009) and the other a velocity-based method due to Vig et al. (2009). In Section 3.2.4 a novel nonparametric method for classifying fixations is introduced and tested against expert classification.

3.2.1 Data

The data used for testing the methods described in this chapter comprised free movements of the eye during a relatively naturalistic search task and were obtained as part of an experiment described in Chapter 4. Only relevant details are given here. Eye movement data were recorded from seven observers viewing twenty natural scenes in blocks of 260 trials, yielding 36400 trials in total, each lasting 1 s. The 86% of trials in which there was no loss in tracking were used in the following.

3.2.2 The method due to van der Linde et al.

The method due to van der Linde et al. (2009) defined a fixation as a sequence of PoG positions which remained within a circle of diameter 1 deg for 100 ms. The thresholds quoted were based on the recommendations by Applied Science Laboratories (Bedford, MA). Figure 3.4 shows the total number of fixations extracted over all of the data analysed when the stability and duration thresholds were varied. Clearly the number of fixations extracted is dependent on the thresholds used. Having a stability threshold which is too large might mean that two distinct fixations are mistakenly classified as one, but having a stability threshold that is too small might mistakenly reject a fixation.
Figure 3.4: The number of fixations produced by the method due to van der Linde et al. (2009) for various stability and duration thresholds.
because of fixational eye movement. This may give the impression that an optimum spatial threshold could be selected as being the threshold which maximizes the number of fixations, thus avoiding these two situations, but a stability-based method requires a duration threshold, which could not be objectively obtained. As shown in Figure 3.4, increasing the duration threshold would simply reject more candidate fixations because they are too short.

### 3.2.3 The method due to Vig et al.

The method due to Vig et al. (2009) was based on the work of Böhme et al. (2006) and used two speed thresholds $v_1$ and $v_2$ to classify saccades. Fixations could be classified as the periods between saccades. Saccade detection was initiated when PoG speed first exceeded the higher threshold of 137.5 deg s$^{-1}$. The onset and offset of the saccade were then defined, respectively, as the first samples where PoG speed rose above and fell below the lower speed threshold of 17.5 deg s$^{-1}$. These thresholds differed a little from those used by Böhme et al. (2006) as they were hand-tuned to match manually classified saccades (personal communication with E. Vig, May 2010). The use of two speed thresholds was intended to increase the noise resilience of this method. Figure 3.5 again shows that the number of fixations extracted is dependent upon the choice of thresholds.
Figure 3.5: The number of fixations produced by the method due to Vig et al. (2009) while varying the stability and duration thresholds.
3.2.4 An objective method for classifying fixations

Given the lack of consensus over acceptable parameters used in many methods for classifying fixations, a more objective method is desirable. Previous attempts to develop data-driven methods for classifying fixations were summarized in Section 3.2. All have, to a greater or less extent, retained parametric dependence. Proposed here is a method which exploited some of the more general distributional properties of eye movements to construct a method for classifying fixations. The method was primarily speed-based, but, by contrast with existing methods, the optimal speed threshold for classifying saccades—and therefore fixations—was derived automatically from the data for each observer and stimulus individually. The method for finding this optimum speed threshold was founded on Tibshirani et al.’s (2001) ‘gap statistic’ for identifying the optimum number of clusters in a set of data. Because speed-based methods can generate unphysiologically short fixations (Nystrom and Holmqvist, 2010), usually attributed to instrumental noise (including noise due to observer movement), the proposed method was extended to include a duration threshold—also inferred from the data—as part of the classification. The accuracy of the proposed method in classifying fixations was assessed by comparison with independent classifications by three experts. The accuracy of two existing parametric methods was measured in the same way, and their performance compared with that of the proposed method.

3.2.5 Classification method

The rationale for the proposed method of classification was that the eye reaches much higher speeds during saccades than during fixational eye movements—such as microsaccades and tremor—and that there are far fewer peaks in speed due to saccades than to fixational eye movements. These two properties of the data made possible the construction of an optimum speed threshold that best separated the speed distributions
of saccades and fixational eye movements. An optimum duration threshold was then constructed that best separated the duration distributions of fixational eye movements and instrumental noise.

**Classification of saccades**

Let $x_i$ and $y_i$ be the horizontal and vertical components respectively of PoG, in screen coordinates, in the $i$th video frame $F_i$. The PoG speed $v_i$ for $F_i$ was defined as the Euclidean distance between the PoG in the immediately preceding and following frames divided by the corresponding difference in time:

$$v_i = \frac{\sqrt{(x_{i+1} - x_{i-1})^2 + (y_{i+1} - y_{i-1})^2}}{|t_{i+1} - t_{i-1}|}. \quad (3.3)$$

This estimate is equivalent to averaging estimates based on the preceding and following successive differences.

A local maximum $v_{max}$ in PoG speed was defined as a PoG speed that was greater than that in the immediately preceding and following frames. By assumption, some of these local maxima were due to saccades whereas others were due to fixational eye movements. The two types of local maxima were separated by a variable speed threshold that was allowed to range between the lowest and highest recorded speeds in the data set in 250 equal steps. The number of steps used here was chosen to reflect the constraints imposed by the temporal resolution of the eye-tracker.

The grey histogram in Figure 3.6 shows, for one observer viewing one image over one block of 223 trials, the distribution of the number of local maxima in PoG speed exceeding a variable speed threshold that ranged between the lowest and highest recorded speed maxima (in this block 37 trials contained lost tracking). For high values of threshold, the local maxima exceeding it were attributed to saccades and, for low values, mainly to fixational eye movements. The much greater number of the latter was
Figure 3.6: Distribution of point-of-gaze speeds. The grey histogram shows the number of local speed maxima exceeding a variable threshold that ranged between the lowest and highest recorded speed maxima. The dashed line shows the number of local speed maxima exceeding threshold under the null distribution, according to which the maxima are uniformly distributed. A red dotted curve marks the gap between the two, but it is hidden by the solid curve, which shows the loess smooth of the same data. Based on 223 trials by one observer viewing one image.
reflected in the rapid increase in the number exceeding threshold for threshold values near zero. The position of the elbow in the histogram was assumed to provide the optimum speed threshold \( v_{opt} \) for classifying saccades.

To locate the elbow, a null distribution of local speed maxima was constructed (Tibshirani et al., 2001) under which values were uniformly distributed. The dashed line in Figure 3.6 shows for this null distribution the number of local maxima exceeding threshold as a function of threshold. This function necessarily declined linearly. The gap between the number of local speed maxima exceeding threshold and the number of null-distribution local maxima exceeding threshold is marked by a red dotted curve in Figure 3.6, but it is hidden by the solid curve, which is explained shortly. The maximum in this gap statistic indicates the location of the elbow and therefore the value of \( v_{opt} \) that optimally separated the distribution of fixational eye movements from the distribution of saccades.

The location of the maximum was estimated from the data by applying a locally weighted quadratic regression (loess) to smooth the histogram. The closeness of the fit was determined by a bandwidth \( h \) which controlled the size of the local neighborhood and consequently the proportion of data included in each local fit (Cleveland, 1979). In general, if the bandwidth is too large, the loess fit is likely to be biased, and with it the location of the maximum; if the bandwidth is too small, the loess fit is likely to incorporate random fluctuations in the data, and therefore produce multiple local maxima. Accordingly, the optimum \( h_{opt} \) was defined as the smallest value of \( h \) for which the number of local maxima \( N \) in the smoothed curve was one, i.e. \( N = 1 \). To accommodate the high gradient at low speed thresholds, speeds were log-transformed before the fit was made.

The solid curve in Figure 3.6 shows the loess smooth and the vertical line indicates the optimum speed threshold \( v_{opt} \). Eye movements whose speed exceeded \( v_{opt} \) were classified as saccades; all other eye movements were classified as fixational or the
CHAPTER 3. COLLECTION AND ANALYSIS OF EYE MOVEMENT DATA

Figure 3.7: Distribution of candidate fixation durations. The grey histogram shows the number of fixation durations and instrumental noise events at very short durations. The solid curve shows the loess smooth of the histogram. Based on the same set of data as in Figure 3.6.

result of instrumental noise.

**Classification of noise**

Figure 3.7 shows for the same set of data as in Figure 3.6 the distribution of durations classified as non-saccadic, that is, for which speed remained continuously less than or equal to the optimum speed threshold $v_{opt}$. There is a local duration maximum at about 250 ms, a value that is consistent with other estimates of typical fixation duration (Castelhano et al., 2009; Henderson, 2003; Rayner, 2009). There is another local duration maximum, which is also a global maximum, at very short durations (less than
a few ms). As mentioned earlier, this maximum was taken to be due to instrument-
tal noise events (Nystrom and Holmqvist, 2010), produced by the observer’s head or
body movement relative to the eye-tracker, or by environmental vibration, or both.
These unphysiological durations have, in other studies, been eliminated by imposing
a minimum duration threshold (Nystrom and Holmqvist, 2010; Tatler et al., 2005). In
keeping with the present nonparametric approach, the number of durations classified
as non-saccadic was assumed to reach its minimum at the value $d_{\text{opt}}$ that optimally sep-
arated the noise distribution from the distribution of fixational eye movements, that is,
where the probability of misclassification was the least and the same for each.

The location of the minimum was estimated from the data by again applying a
quadratic loess to smooth the histogram. As with the speed data, the optimum value
$h_{\text{opt}}$ of the bandwidth was determined nonparametrically. Thus, the optimum $h_{\text{opt}}$ was
defined as the smallest value of $h$ for which the number of maxima $N$ in the smoothed
curve was two, i.e. $N=2$. To accommodate the high gradient at low durations, speeds
were log-transformed before the fit was made.

The solid curve in Figure 3.7 shows the loess smooth and the vertical line indicates
the optimum duration threshold $d_{\text{opt}}$.

**Comparison with expertly classified fixations**

The proposed method was applied to the sets of data described in Section 3.2.1. Esti-
mates of the optimum speed threshold $v_{\text{opt}}$ and duration threshold $d_{\text{opt}}$ were estimated
independently and automatically for each observer viewing each image over each block
of 260 trials (excluding, as noted earlier, trials with any lost tracking). The method
could just as well have been applied to trials pooled over blocks, sessions, or scenes.
The light-grey histograms in Figure 3.8 show the distribution of optimum speed and
duration thresholds for the 20 scenes and seven observers tested.
Figure 3.8: Distribution of optimum speed thresholds $v_{opt}$ and optimum duration thresholds $d_{opt}$ over 140 blocks of 31409 trials (light-grey histograms). Also shown is the distribution of $v_{opt}$ and $d_{opt}$ for blocks from which a subsample of 21 trials was taken from the 31409 trials for expert classification (dark-grey histograms).
A subsample of the resulting fixation classifications was compared with hand-classifications made independently by three experts, RA, GRB, and EG\(^1\). Twenty one PoG traces were selected at random from the 31409 available (three each from six of the observers, and two from the seventh). All 21 traces were then classified by RA, 18 traces by EG, and three traces by GRB. The agreement between classifications by the proposed method and those by the three experts was summarized by the proportion of frames with common classification. The total number of frames in 21 traces was 5271. Results are shown in Table 3.1, with classification accuracy ranging from 88% to 94%. An example of a trace with fixations classified by the proposed method and by one of the experts is shown in Figure 3.9. A two-dimensional representation of the same data data is shown in Figure 3.10.

To help set the performance of the proposed method in context, the classification accuracy of the two parametric methods described in Sections 3.2.2 and 3.2.3, one based on eye stability, the other on speed, was also assessed against the classifications by the experts. As with the proposed method, the agreement between the two parametric methods and the classifications by the experts was summarized by the proportion of frames with common classification. Results are again shown in Table 3.1. Classification accuracy ranged from 85% to 96%.

To provide an estimate of the best performance that could be expected from a fixation classification method, the pairwise agreement between the experts was summarized by the proportion of commonly classified frames. The mean agreement over the 24 comparisons was 93%. Since expert classification provides the best available estimate of a ground truth, this value sets an upper limit on average classification accuracy.

\(^1\)The three independent expert classifiers were Dr. R. Ackerley (Sahlgrenska Hospital and Institute for Neuroscience & Physiology, University of Gothenburg, Sweden), Professor G. R. Barnes (Faculty of Life Sciences, University of Manchester, UK) and Dr. E. Gowen (Faculty of Life Sciences, University of Manchester, UK).
Figure 3.9: An example of a typical eye movement trace. Vertical and horizontal gaze locations are plotted as a function of time from the beginning of the sample. Fixations classified by the proposed method (upper plot) and by expert EG (lower plot) are indicated by thicker curves. The classification agreement between these two plots was 95%. A two-dimensional representation of these data is shown in Figure 3.10.
Figure 3.10: A 2 dimensional plot of the eye movement trace shown in Figure 3.9. Fixations classified by the proposed method (upper plot) and by expert EG (lower plot) are indicated by the red circles.
Table 3.1: Agreement (%) between fixation classifications by each experimental method and by three experts across 21 sample traces (5271 frames). Expert RA classified all 21 traces; expert EG classified 18 traces (4518 frames); expert GRB classified 3 traces (753 frames).

<table>
<thead>
<tr>
<th>Experimental classifier</th>
<th>Expert classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RA</td>
</tr>
<tr>
<td>Proposed nonparametric method</td>
<td>94</td>
</tr>
<tr>
<td>Vig et al. (2009) method</td>
<td>96</td>
</tr>
<tr>
<td>van der Linde et al. (2009) method</td>
<td>92</td>
</tr>
</tbody>
</table>

Robustness to added noise

One of the advantages of a classification method that derives thresholds automatically from data over a method that uses a fixed threshold is its adaptability to different levels of instrumental noise (including that due to observer movement). If data are particularly noisy, a speed-based classification method may misrepresent noise events as saccades. Misclassifications may be reduced by increasing the value of the speed threshold after visual inspection of the data, but these adjustments can be uncertain and time consuming, particularly if noise levels vary across observers.

To demonstrate the robustness of the proposed method, a simulation was undertaken in which different levels of Gaussian noise were added to the data before analyzing by the proposed method. Figure 3.11 shows for each noise level $\sigma$, the automatically generated optimum speed threshold $v_{\text{opt}}$ averaged over all 140 blocks (upper panel). The threshold $v_{\text{opt}}$ increases smoothly with increasing $\sigma$. Crucially, the total number of fixations classified by the method (lower panel) remains almost constant with $\sigma$. 
Figure 3.11: Effect of added noise. Optimum speed threshold $v_{opt}$ averaged over all 140 blocks (upper panel) and corresponding number of fixations classified by the proposed method (lower panel) are plotted against added noise $\sigma$ in degrees of visual angle.
3.3 Discussion

The work described in subsequent chapters of this thesis contains conclusions which were informed by eye movement data, so the techniques and equipment used for the collection and processing of such data were important. The eye-tracker used in this project was tested with an artificial eye to obtain estimates of error in the device. The error was estimated to be approximately 0.2 deg. Testing was also undertaken using human observers to include other errors which might be encountered in practice, such as those attributable to involuntary movements of the observer’s eyes or head. The error estimated using human observers was approximately 0.5 deg.

Three different approaches to calibration of a human observer were tested, two of which were intended to account for drift of an observer over the course of an experimental block. It was found that there was no noticeable advantage in attempting to correct data for drift. The details of the calibration procedure were improved over the course of the project, and so calibration will be described for each experiment individually.

In this chapter the problem of the classification of fixations was addressed. As was shown in Sections 3.2.2 and 3.2.3, the fixations extracted by different methods can be influenced by the values used for speed, duration or stability thresholds. Also, selecting the values to be used for thresholds is not an easy task, given inter-observer differences, and the effect of scene content and task. The proposed nonparametric method for classifying fixations solves such problems, as it automatically adapts itself to such effects. Unlike the other methods discussed, it requires no subjective hand-tuning of thresholds and therefore no intervention by an independent eye movement expert. Because it is automatic, it can be applied to large quantities of data, with corresponding gains in classification rate and subsequent analysis.
Although the approach adopted in this method was based on the general distribu-
tional properties of eye movements, only the existence of extrema in the distributions
was assumed, i.e. a maximum or a minimum. No particular assumption was made
about the model of the distribution or its parameters. As with any data-driven proce-
dure, however, performance necessarily depends on the size of the sample. With the
samples of 260 trials used here, extreme estimates of the speed and duration thresh-
olds demarcating fixations were only rarely encountered, as Figure 3.8 made clear. Of
course, smaller samples would lead to greater variability in individual fixation classi-
fications.

The proposed method was found to be successful in that it produced fixation clas-
sifications that agreed with 88-95% of the independent classifications by the three ex-
erts. This level of agreement was generally as good as—and for some expert classi-
fications better than—the level with the parametric methods due to Vig et al. (2009)
and to van der Linde et al. (2009). Significantly, none of the methods of classification
agreed perfectly with the experts, but given the level of agreement between the experts,
some differences were inevitable. The proposed nonparametric method will be used in
analyses performed in the remainder of this thesis.
Chapter 4

Searching Images of Natural Scenes

As was discussed in Section 1.5, many studies of visual search have investigated search in artificial search arrays rather than natural scenes, which may be an oversimplification of real-world search behaviour. Motivation for using artificial search arrays include the ease of control and repeatability. Unlike natural scenes, unlimited numbers of artificial search arrays with similar visual properties can be generated, which is useful because these experiments generally require that similar stimuli are presented many times (Wolfe, 1994b). Perhaps inspired by the proliferation of eye-tracking equipment, many recent studies discussing search in natural scenes have concentrated on the dynamics of eye movements rather than the image properties influencing search difficulty (Torralba et al., 2006; Tatler, 2007; Einhäuser, Rutishauser and Koch, 2008; Zhang et al., 2008; Castelhano et al., 2009). The varied success of attempts to predict eye movements was discussed in Section 1.6. The aim of this chapter is to address the question of which image properties affect performance when observers are searching natural scenes.

When applying the Feature-Integration Model of Attention (Treisman and Gelade, 1980), introduced in Section 1.5, to artificial search arrays with discrete targets and distractors one can usually decide whether the search can be performed in parallel (a
Figure 4.1: A road sign warning that cycling is not allowed (Herbert, 2010).

Figure 4.2: A lizard on some grass (Heffner, 2009).
feature search) or whether search must be achieved by serially inspecting each element (a conjunction search). With a conjunction search, task difficulty can be predicted as a function of distractor density. The Feature-Integration Model of Attention can broadly be applied to search with images of natural scenes. For example, it is clear that the road sign in Figure 4.1 differs from the background along one feature direction: colour. The sign ‘pops out’ and search for the sign could be performed in parallel. An example of a conjunction search would be to search for the lizard in Figure 4.2. The difficulty of a conjunction search in natural scenes is not easy to predict because we cannot simply measure the density of discrete distractors.

Rosenholtz et al. (2007) suggested that search difficulty could be predicted for complex imagery by substituting measures of visual clutter for the distractor density (or ‘set size’) used in search arrays. They tested their clutter measures on geographic maps and their work was extended to images of natural scenes by Henderson et al. (2009). In both studies three measures of clutter were used: feature congestion, sub-band entropy and edge density (these measures are explained in more detail in Sections 4.2.6, 4.2.7 and 4.2.8). The targets used by Rosenholtz et al. (2007) were Gabor filters and arrow symbols; Henderson et al. (2009) used small Ts and Ls embedded in images of natural scenes which observers were asked to discriminate. Search performance was measured by Rosenholtz et al. (2007) using reaction time and by Henderson et al. (2009) using both reaction time and search failure. Both studies found that each measure of clutter was correlated with search performance measures. Henderson et al. (2009) noted that although clutter accounted for a statistically significant amount of variance in search performance, it was a relatively small amount compared with the influence of set size observed in visual search tasks using artificial displays with discrete targets and distractors. Although one of Rosenholtz et al.’s (2007) clutter measures did include a measure of colour variability, neither study tested the influence of more basic image properties upon search performance. They also looked at global measures of clutter
rather than properties local to the target location.

In the experiment described here a search task was undertaken by observers and detection performance was measured at 130 locations on each of 20 images of natural scenes, allowing the production of twenty ‘detection profiles’ of resolution 13 \times 10. The target was a small grey target superimposed digitally on the image. The target was contextually neutral (that is, there was no particular location within the image where it was more or less likely to be found), so observers would not exhibit biases due to scene gist and context (Torralba et al., 2006), and it was chromatically neutral. Detection performance was quantified by $d'$ (explained in Section 4.2). Owing to the large number of trials required, trial durations were kept short. Observers’ gaze was monitored throughout each trial and observers were shown each image before they began the trials to reduce the effects of novel stimulus onset. The proportion of variance in detection performance accounted for by a range of image properties was calculated. Image features analyzed included Rosenholtz et al.’s (2007) proposed clutter measures.

4.1 Method

4.1.1 Apparatus

Gaze was monitored with a High Speed Video Eyetracker Toolbox (250Hz; Cambridge Research Systems Ltd, UK) described in Chapter 3. Images were displayed on a 20-inch RGB CRT colour monitor (GDM-F520, Sony Corp., Tokyo, Japan) controlled by a graphics workstation (Fuel V12, Silicon Graphics Inc. Mountain View, CA). The spatial resolution of the display was 1600 \times 1200 pixels; the intensity resolution on each RGB gun was 10 bits; and the refresh rate was approx. 60 Hz. The display subtended approximately 17 deg \times 13 deg at a viewing distance of 1m.
CHAPTER 4. SEARCHING IMAGES OF NATURAL SCENES

4.1.2 Stimuli

Twenty colour images of natural urban and rural scenes, taken from a set of hyperspectral images (Foster et al., 2004), were used in the experiment. The mean luminance of the images on the screen was 3.6 cd m$^{-2}$ (range 0 – 61.4 cd m$^{-2}$). The target was a neutral grey sphere subtending approximately 0.3 deg (Munsell N7) superimposed on the image and matched in luminance to its local surround. The target was designed so as to be subtle, but detectable, and was grey to be chromatically neutral. It was also important that the target had no relation to scene context as the aim of this experiment was to investigate image properties rather than target properties or the interaction of target properties with scene properties (Torralba et al., 2006). When present, the target was randomly positioned within one cell of a 10 × 13 grid. It appeared once in each cell for each subject, and was absent in half of the trials. Some examples of the stimuli are shown in Figure 4.3.

4.1.3 Observers

Seven observers (3 male, 4 female, aged 21 to 31 years) took part in the experiment. All had normal colour vision and normal or corrected-to-normal visual acuity. All except one of the observers were unaware of the purpose of the experiment.

4.1.4 Experimental procedure

The observer was seated with his or her head held steady by chin and forehead rests in front of a CRT monitor on which images were presented. In each trial, an image with or without the target was presented for 1 s, and the observer then indicated whether or not the target was there. Trials were performed in blocks of 260, all containing the same scene, but divided into four sub-blocks of 65 trials separated by a short break.
Figure 4.3: Some examples of the images used in this experiment. The target is present in the top right image and is indicated with the red arrow.
Each sub-block lasted 5–10 min and one block was performed in each session of approximately 1 hr.

Calibration data for the eye-tracker were recorded at the start, middle, and end of each sub-block, as follows. Twenty calibration targets, arranged in a $5 \times 4$ grid, were presented on the screen at known positions. Observers were asked to click a mouse button as they fixated each target, and the corresponding horizontal and vertical PoG signals were recorded by the eye-tracker. This produced three sequences of 20 measured gaze positions, $((x_{1,1}, y_{1,1}), ..., (x_{1,20}, y_{1,20})), ((x_{2,1}, y_{2,1}), ..., (x_{2,20}, y_{2,20})), \text{ and } ((x_{3,1}, y_{3,1}), ..., (x_{3,20}, y_{3,20}))$. If for any of the three sequences a valid PoG signal for a target was unavailable owing to a loss in tracking, it was replaced by the corresponding PoG signal from another sequence (at least one replacement was necessary in 55% of sub-blocks). For the first half of the sub-block, the sequences $((x_{1,1}, y_{1,1}), ..., (x_{1,20}, y_{1,20}))$ and $((x_{2,1}, y_{2,1}), ..., (x_{2,20}, y_{2,20}))$ were averaged point-wise and fitted to the target positions, expressed in screen coordinates, by a linear transformation to give least-squares error. This transformation was used to transform to screen coordinates all the experimental PoG signals obtained between this first and second calibration. An analogous procedure was used to transform to screen coordinates all the experimental PoG signals obtained between the second and third calibration.

For each of these calibration transformations, the RMS difference between the calibration targets and the position of the observer’s gaze on the screen during fixation of the target provided an estimate of the calibration error (yielding two such error estimates for each sub-block). The mean value of the calibration error was 0.26 deg and individual estimates did not exceed 0.51 deg.

In total, data were collected from 5200 trials for each of the seven observers participating in this experiment.
CHAPTER 4. SEARCHING IMAGES OF NATURAL SCENES

4.2 Results and analysis

Detection ability was quantified in this experiment using $d'$ from signal detection theory (DeCarlo, 1998). This value is calculated as follows

$$d' = z(H) - z(F),$$

(4.1)

where $H$ is the hit rate, $F$ is the false-alarm rate, and $z$ is the inverse of the cumulative normal distribution function. The advantage of calculating $d'$ rather than hit rate alone is that $d'$ accounts for observer bias towards a ‘yes’ or ‘no’ response. For each image, $d'$ was calculated for each of the 130 cells over all seven observers, producing a detection profile for each image. Figure 4.4 shows the profiles for the images shown in Figure 4.3.

There are a number of image properties which were considered as possible contributors to detection ability:
• fixation density.

• luminance.

• chroma.

• luminance entropy.

• luminance contrast.

• edge density.

• clutter.

It should be noted that these image properties may not necessarily be independent; for example luminance contrast is likely to be be correlated with edges (Baddeley and Tatler, 2006) and the perceived chromatic difference between a target and its background may be correlated with perceived luminance contrast (Walkey et al., 2005).

For each image, a 13 × 10 profile of each of these properties (luminance, chroma, etc.) was produced. In the first stage of analysis, the coefficient of determination $R^2$ between each property profile and the detection profile was calculated for each image. The mean was taken of the resulting twenty values for each property. In addition to these image properties, fixation density was also considered.

The proportion of variance accounted for by a particular image property could be confounded by the fixation distributions. For example, it may be that a particular image property appears to account for a large proportion of variance in $d'$, when in fact the property was simply correlated with fixation density which may in turn be correlated with detection performance. To take this into account, a second stage of analysis was made. The variance accounted for by a multiple linear regression—which combined the fixation distribution and each image property—was also calculated. The structure
of the model was
\[ d' = aF + bI + cFI, \]  
(4.2)

where \( F \) is fixation density and \( I \) is the value of the image property in each cell. \( a, b \) and \( c \) are constants which are selected to optimize the fit.

Fixations were extracted using the method described in Section 3.2. Dropped frames were removed from the data before fixations were extracted and accounted for 2.08\% of the data. The fixation density map was created by binning fixations into a \( 13 \times 10 \) histogram and normalizing the result. Fixations alone accounted for 28\% of variance in detection ability. Details of the image property profiles are given in the following sections.

### 4.2.1 Luminance profile

The luminance profile was created by simply downsampling the luminance map to a resolution of \( 13 \times 10 \). Luminance accounted for 20\% of variance in detection ability.

### 4.2.2 Chroma profile

It was necessary to convert the colour information of the images from CIE 1931 XYZ colour space to CIELAB before producing the chroma map (see Section 1.8). As with the luminance profile, the \( 13 \times 10 \) chroma profile was produced by simply downsizing the full resolution chroma profile. Chroma accounted for 29\% of the variance in detection ability.

### 4.2.3 Luminance entropy profile

The \( 13 \times 10 \) luminance entropy profile was produced by estimating the luminance entropy in each cell of the grid using Kozachenko and Leonenko’s (1987) estimator,
using a program implemented by Iván Marín-Franch. This accounted for 13% of the variance in detection ability.

### 4.2.4 Luminance contrast profile

The $13 \times 10$ luminance contrast profile was produced by estimating the luminance contrast in each cell of the grid, defined here as the standard deviation of the luminance values. This measure of contrast is sometimes known as RMS contrast (Peli, 1990). It accounted for 11% of the variance in detection ability.

### 4.2.5 Edge density profile produced using Gabor filters

Edge density content was extracted by convolving each of the images with oriented Gabor filters (Tatler et al., 2005). The images were symmetrically padded with a border equal to the width of the filter before convolution. The Gabor filter is composed of a sinusoidal ‘carrier’ function modulated by a Gaussian ‘envelope’ function and can be written (for a vertically oriented filter) as follows:

$$f(x, y) = \exp \left(-\frac{(x^2 + y^2)}{2\sigma^2}\right) \sin \left(\frac{2\pi x}{\lambda}\right),$$

where $\sigma$ is the standard deviation of the Gaussian envelope and $\lambda$ is the wavelength of the carrier. An example is shown in Figure 4.5.

Four oriented edge density maps were produced, with the filter oriented at 0, $\pi/4$, $\pi/2$ and $3\pi/4$ radians. The maps were then combined by taking the maximum value of the four orientations for each pixel in the image (Tatler et al., 2005). The wavelength of the sinusoidal carrier $\lambda$ was fixed at $\lambda = 18$ pixels (equivalent to carrier frequency $f_c = 4$ cycles per degree (cpd)). The standard deviation of the Gaussian envelope $\sigma$ was set to $2.5\lambda$, and the size of the filter was set to $6\sigma$. Because increasing the wavelength of the carrier, and therefore the size of filter, was impractical owing to the
increased run times, edge density for different scales was measured by retaining the parameters of the filter, but reducing the resolution of the image before convolution. Images were scaled by factors of 0.5, 0.25 and 0.125 before being convolved with the Gabor filter. This was equivalent to convolving the full resolution images with Gabor filters of $f_c = 2, 1$ and 0.5 cpd or $\lambda = 36, 72$ and 144 pixels (the width of the target was 22 pixels). After convolution each edge density map was downsized to $13 \times 10$ pixels. Edge density accounted for little variance in $d'$; the greatest amount of variance by edge density at the four different scales was 8%.

### 4.2.6 Edge density profile produced using Canny edge detector

The method for calculating edge density used by Rosenholtz et al. (2007) and Henderson et al. (2009) was replicated here. First a Canny edge detector was applied to the image. The edge detector requires three parameters: a low threshold, a high threshold and the standard deviation of the Gaussian filter used by the edge detector. The low and high threshold are used to detect weak and strong edges respectively, and weak edges are kept only if they are connected to strong edges. The low and high thresholds were set to 0.11 and 0.27, respectively, and the standard deviation of the Gaussian filter was set to the default value of 1 (Rosenholtz et al., 2007). Edge density was calculated for each cell in the $13 \times 10$ grid as the proportion of pixels in the cell which were...
edge pixels. Edge density calculated in this way accounted for 9% of the variance in detection ability.

4.2.7 Feature congestion profile

Rosenholtz et al. (2007) based the measure of feature congestion on the analogy that the more cluttered a display or a scene is, the more difficult it would be to add a new item that would reliably draw attention. They extracted luminance contrast, colour and orientation at a range of scales. The local covariance of each feature was calculated at three scales. The resulting three maps were combined across scales by taking the maximum measure at each pixel to form one clutter map for each feature. Finally, the individual clutter maps were combined to form one clutter map. The cube root was taken of the color clutter (as it is a volume) and the square root of orientation clutter (as it is an area) to make the values comparable to contrast congestion (a one-dimensional magnitude). Each clutter map was normalized by its standard deviation before the three were summed to form the final clutter map. Rosenholtz et al. (2007) have provided software for the calculation of feature congestion which was used to create clutter maps for each image. The 13 × 10 feature congestion profile was produced by downsizing the full resolution feature congestion map. Feature congestion accounted for 7% of the variance in detection ability.

4.2.8 Subband entropy profile

Subband entropy (Rosenholtz et al., 2007) was based on Shannon entropy (see Section 1.9). Rosenholtz et al.’s (2007) algorithm used a steerable pyramid, which is a utility for multi-scale, multi-orientation decomposition of an image (Simoncelli and Freeman, 1995). In brief, the algorithm used to estimate subband entropy is as follows:

1. Convert the image into CIELAB.
2. Decompose the luminance and chrominance bands into wavelet subbands using a steerable pyramid Simoncelli and Freeman (1995).

3. Bin the wavelet coefficients within each subband and compute the Shannon entropy within each subband (using Equation 1.8).

4. Sum the subband entropies for the luminance and chrominance channels.

5. Compute a weighted sum of chrominance and luminance entropies. Rosenholtz et al. (2007) used a weighting of 0.08 for each of the chrominance channels and 0.84 for the luminance channels, and pointed out that the measure was not sensitive to changes in the weighting.

As with feature congestion, subband entropy was estimated using software provided by Rosenholtz et al. (2007). A $13 \times 10$ subband entropy map was made by estimating subband entropy in each of the 130 grid cells. Subband entropy accounted for 7% of the variance in detection ability.

### 4.2.9 Variance accounted for by each feature

The proportion of variance of detection ability accounted for by each of the features, averaged over the twenty images, was given at the end of each relevant sections. The values are summarized in Figure 4.6. Chroma accounted for the largest proportion of variance in detection ability, slightly more variance than fixation density. The proportion of variance accounted for by the linear model of fixation density and each image feature (Equation 4.2) is shown in Figure 4.7. In general, creating a linear model of fixation density and each image feature increased the proportion of variance for which the model accounted. The linear model of fixation density and chroma was the strongest predictor of detection performance, and accounted for more variance than either quantity alone. This showed that, rather than the prediction power of chroma
being attributable simply to its correlation with fixation density, chroma was an important factor in the detectability of the target. After chroma, luminance was shown to be the next strongest predictor of detection performance.
Figure 4.6: The proportion of variance of detection performance accounted for by each feature alone. Error bars represent one standard error.
Figure 4.7: The proportion of variance of detection performance accounted by a linear model of fixation density $F$ and each image feature $I$. The model had the form $d' = aF + bI + cFI$, where $a$, $b$ and $c$ are constants which were selected to optimize the fit. The black vertical line shows the proportion of variance of detection performance accounted for by fixation density alone.
4.3 Discussion

The mechanisms of visual search have been subject to considerable attention in the vision community for years, partly because if we are able to expand our understanding of human visual search it may improve attempts to produce computer programs to emulate it. Many previous studies have concentrated on the movements of the eyes during visual search, attempting to predict the motion of the eye over visual stimuli. Although many of these studies have been able to account for at least some eye movements, attempts to predict eye movements will always be limited by the complexity of the processes by which we decide where to look (in particular, uncertainty about the relative contribution of bottom-up and top-down factors).

In the study described here, the contribution of image features to detection performance was investigated. To achieve this, detection performance was measured at each of 130 locations uniformly distributed across the image, and a detection profile was produced for each image. This made it possible to investigate the effect of local image properties on detection performance. It was shown that of the range of features tested, the greatest proportion of variance in detection ability was accounted for by chroma. This result was maintained when image features were combined with fixation density using a linear model.

Interestingly, texture measures (edges, clutter, entropy and contrast) were relatively weak in accounting for variance in detection ability. The various measures of edge density and the clutter measures proposed by Rosenholtz et al. (2007) accounted for little variation in detection performance. This result does not necessarily contradict the findings of Rosenholtz et al. (2007) and Henderson et al. (2009) as there were a number of differences between the studies. Here image features were measured locally rather than globally. There were also differences in how images were presented. In this study each observer saw each image in a total of 260 trials, far more than in the studies of
either Rosenholtz et al. (2007) and Henderson et al. (2009). The large number of trials was necessary to build the full detection profile for each image, but may have allowed observers sufficient time to develop search strategies.

The fact that chroma was shown to be so important in accounting for variance in detection ability might be argued to be due to the nature of the target used. The grey sphere was an extremely simple target in terms of structure and so perhaps was less likely to be lost in regions of high edge density, unlike the Gabor patches and arrows used by Rosenholtz et al. (2007) and Ts and Ls used by Henderson et al. (2009), and would be likely to be more conspicuous when superimposed on a region of bright colour than on a grey region.

Luminance was a strong predictor of detection performance, accounting for the third largest proportion of variance in detection ability after chroma and fixations and the second largest proportion of variance in detection ability when fixations were taken into account. This may seem surprising since the target was matched in luminance to its local surround. However, it is important to note that the target was a sphere with shading, not an isoluminant disc, so would never exactly match the background luminance.

It is interesting that what are arguably the two most basic image features investigated accounted for the most variance in detection ability in this study. Higher order features such as entropy and contrast do still account for variance in detection ability, although to a lesser extent. This finding bears similarities to the findings of a study by Açı̇k et al. (2009) on fixation allocations. They concluded that high- and mid-level features, such as edges, will guide fixations where they are present, but otherwise low-level features prevail. The results of this experiment seem to support this statement. The study described in this chapter used a range of images, but it could be argued that an image of a city scene would have broadly similar high-level properties across
the image, and likewise for rural scenes. Perhaps if the experiment were repeated using images with a larger variation in high-level features these features would account for detection performance. Other further work could include testing the generality of the results here by reproducing the experiment, but altering the nature of the target, perhaps to a high-chroma target.
Chapter 5

Steering Human Fixations

Whenever we gaze along a busy street trying to locate the friend we’re due to be meeting, only to find our gaze falling upon a brightly coloured advertisement for toothpaste, an advertising executive has succeeding steering our visual attention, albeit rather un-subtly. Steering attention is also the aim of the Computer-Aided Diagnosis programs used in medical imaging to direct radiologists to potential tumours (Doi, 2007). More speculatively, gains could be made in image compression if attention could be reliably steered away from unimportant regions. Previous attempts to develop algorithms to systematically steer the fixations of an observer by altering image properties were summarized in Section 1.6.

As was discussed in Section 1.7.6, edges have often been linked to human attention. The experiments in this section used images in which regions were modified using unsharp marking, which was used to decrease or increase the sharpness of edges in particular image regions. The first approach aimed to bias fixations away from a particular image region using an algorithm which reduced sharpness, but this objective was not achieved. Possible reasons for this are discussed (see Section 5.1). The second approach aimed to attract fixations towards a region by sharpening local image regions. This objective was achieved, and the modification also improved search performance
(see Section 5.2).

5.1 Smoothly decreasing de-enhancement

The images used in the experiments detailed in this section were manipulated by reducing their edge sharpness (see Section 1.7.6). The modification, known hereafter as de-enhancement, was at its maximum level at the right or left side of the image, decreasing smoothly to no de-enhancement at the opposite edge and will be referred to in the following as ramp de-enhancement. A smooth change was used rather than a step change to reduce artifacts. The aim of de-enhancement was to make regions of the image less visually interesting, and thus reduce the probability of them being fixated. Three experiments were conducted using de-enhancement and are listed below.

1. **Ramp de-enhancement with target-search task.** Observers were asked to search images for the target. Images were either unmodified or had been de-enhanced to the left or the right.

2. **Ramp de-enhancement with change-detection.** Observers were presented with an image twice in each trial. The two presentations showed the image either under the same condition twice or under two different conditions. Observers were asked to indicate whether they had detected any change.

3. **Ramp de-enhancement with forced choice target-search task.** Observers were presented with an image twice in each trial. The two presentations showed the image either under the same condition twice or under two different conditions. A target was superimposed on the image in the first or second presentation. Observers were asked to indicate whether the target was superimposed on the image in the first or second presentation.
5.1.1 Stimuli

In these three pilot experiments, six of the twenty colour images of natural urban and rural scenes used in the experiment described in Chapter 4 were used. An example of a de-enhanced image is shown in Figure 5.1 and the other five images are shown in Figure 5.2. Where it was used, the target was the same as that described in Chapter 4, namely a neutral grey sphere (Munsell N7) subtending approximately 0.3 deg superimposed on the image and matched in luminance to its local surround.
Figure 5.2: Images used in de-enhancement experiments.
5.1.2 Apparatus

Apparatus used in these pilot experiments was the same as that used in the experiment described in Chapter 4.

5.1.3 Ramp de-enhancement with target-search task

Observers

Five observers took part in the experiment (all female, aged 20 to 22 years). All had normal colour vision and normal or corrected-to-normal visual acuity. Observers were not informed of the purpose of the experiment. Two of the observers had also taken part in the experiment described in Chapter 4.

Procedure

Observers were seated with their head held steady by a forehead rest and chinrest in front of a CRT monitor on which images were presented. In each trial observers viewed an image of a natural scene for 1 s while searching for the target with free eye movements. Observers indicated by a mouse click when they were ready to begin the session and after each trial indicated, again by a mouse click, whether or not the target had been detected. There was no limit in their response time. The probability of a target being present in a trial was 0.5. Trials were grouped into blocks consisting of sixty trials. The images used in the first thirty trials could either have no de-enhancement (condition N1), the left side de-enhanced (condition L1) or the right side de-enhanced (condition R1), and the images used in the second thirty trials could also be presented with any of the three image conditions (condition N2, condition L2 and condition R2). Observers performed 840 – 1140 trials each.

The calibration procedure was the same as that in Section 4.1.4. Calibration took place at the beginning, middle and end of each block, and if valid point-of-gaze (PoG)
signals for any of the target positions were unavailable owing to a loss in tracking, they were replaced by the corresponding PoG signals from another calibration (this was necessary in 16% of blocks). If the calibration error (estimated, as before, by calculating the RMS distance between the calibration targets and the position of the observer’s PoG on the screen during fixation of the target) exceeded 0.30 deg the data from the block was not included in analysis and the block was repeated. The mean calibration error for blocks included in analysis was 0.22 deg.

Results and analysis

The method described in Section 3.2.4 was used to extract fixations from the PoG data. Dropped frames, which accounted for 0.76% of the data in this experiment, were removed from the data for each trial before fixations were extracted. Speed and duration thresholds were derived individually for each observer viewing each image. Only fixations after the first saccade in each trial were included in analysis.

To test the effect of the enhancement on fixations, the horizontal components of fixations for each image were pooled over all observers for condition N1 and condition N2 trials to form a vector $F_N$. Likewise the horizontal component of fixations was pooled over all observers for condition L1 and condition L2 trials to form a vector $F_L$ and the horizontal component of fixations was pooled over all observers for condition R1 and condition R2 trials to form a vector $F_R$. The mean of each of these vectors was calculated. The results are shown in Figure 5.3. A factorial ANOVA (image $\times$ condition) revealed a significant effect of image on the mean of horizontal fixation location ($F(5, 10) = 20.04$, $p < 0.001$), but no significant effect of condition ($F(2, 10) = 3.46$, $p=0.072$).

The effect of de-enhancement was also tested using KL divergence (introduced in Chapter 2). The horizontal component of fixations was pooled within each image over all observers for condition N1 trials to form a vector $F_{N1}$, condition N2 trials to form a
Figure 5.3: The mean of horizontal fixation locations under each condition (no de-enhancement, left side de-enhanced and right side de-enhanced) for ramp de-enhancement with target-search task.
vector $F_{N2}$, condition L1 trials to form a vector $F_{L1}$, condition L2 trials to form a vector $F_{L2}$, condition R1 trials to form a vector $F_{R1}$ and condition R2 trials to form a vector $F_{R2}$. KL divergence was calculated between the horizontal components of fixations from trials under the same conditions ($d_{KL}(F_{N1}, F_{N2})$, $d_{KL}(F_{R1}, F_{R2})$ and $d_{KL}(F_{R1}, F_{R2})$) and between the horizontal components of fixations from trials under different conditions ($d_{KL}(F_{N1}, F_{L2})$, $d_{KL}(F_{N1}, F_{R2})$, $d_{KL}(F_{R1}, F_{L2})$, $d_{KL}(F_{R1}, F_{N2})$, $d_{KL}(F_{L1}, F_{N2})$ and $d_{KL}(F_{L1}, F_{R2})$). Results are shown in Figure 5.4. KL divergence for was not significantly higher for different conditions than for same conditions (two-sample one-tailed $t$-test, $p = 0.603$).
5.1.4 Ramp de-enhancement with change-detection

The fact that image modification did not have a detectable effect on fixations in the experiment described in Section 5.1.3 may have been because images were presented under the same condition multiple consecutive times so observers might have learned to ignore the modification. The lack of effect could also have been attributed to the use of a search task, which had previously been shown by Einhäuser, Rutishauser and Koch (2008) to override attempts to steer visual attention by image modification. In this experiment the task was altered from a target detection task to a change-detection task. In order to accommodate the change in task, presentation structure was also altered so that observers were presented in each trial with the same image twice. The two presentations showed the image either under the same condition twice or under two different conditions.

Observers

Fifteen observers (5 male, 10 female, aged 19 to 42 years) took part in this experiment. All had normal colour vision and normal or corrected-to-normal visual acuity. Observers were not informed of the purpose of the experiment. Two of the observers had also taken part in the experiment described in Chapter 4 and three had taken part in the experiment described in Section 5.1.3.

Procedure

Again, observers were seated with their head held steady by a forehead rest and chin-rest in front of a CRT monitor on which images were presented. In each trial, observers viewed the same image in two consecutive presentations. Each presentation lasted 1 s and the two were separated by a 200 ms blank period. The blank period was included as showing the two images with no blank period would add a temporal
aspect to the de-enhancement (a ‘sudden change’) that was not present in the previous experiment. The two presentations showed the image either under the same condition twice or under two different conditions. Observers indicated after each trial whether the two images were the same or different. There were five trial types: no de-enhancement followed by no de-enhancement (condition NN), no de-enhancement followed by right side de-enhanced (condition NR), right side de-enhanced followed by no de-enhancement (condition RN), no de-enhancement followed by left side de-enhanced (condition NL) and left side de-enhanced followed by no de-enhancement (condition LN). Each observer saw each image in one block of thirty trials. One third of the trials were condition NN, one third were either condition NR or condition RN and one third were either condition NL or condition LN.

The eye-tracker was calibrated in a manner similar to that in Section 4.1.4, but since fewer trials were undertaken in each block in this experiment, calibration data were recorded only twice during each block, once at the end and once at the beginning. As in the experiment described in Section 4.1.4, if valid PoG signals for any of the target positions were unavailable in either of the two calibrations owing to a loss in tracking, they were replaced by the corresponding PoG signals from the other calibration (this was necessary in 21% of blocks). As before, if the calibration error exceeded 0.30 deg the data from the block was not included in analysis and the block was repeated. The mean calibration error for blocks included in analysis was 0.22 deg.

Results and analysis

As before, the method described in Section 3.2.4 was used to extract fixations from the PoG data. Dropped frames, which accounted for 1.76% of the data in this experiment, were removed from the data for each trial before fixations were extracted. Speed and duration thresholds were derived individually for each observer viewing each image. Only fixations after the first saccade in each trial half were included in analysis.
For each image, the horizontal components of fixation locations were pooled over observers and trials within each scene, condition and trial half (for example, over the first half of all condition NN trials in which the image viewed was ‘ctrunks1b’) to give, for each image and condition, two vectors $F_{h1}$ and $F_{h2}$.

The mean horizontal position of the fixations was calculated for the first and second half of each trial giving $F_{h1}$ and $F_{h2}$ respectively. The difference between the two was calculated as $(F_{h2} - F_{h1})$. If the difference were positive, it would show that fixations in the second half the trial were more biased to the right than the first half. If the effect of de-enhancement were to bias fixations away from the image region to which it was applied, then the value of $(F_{h2} - F_{h1})$ for condition NL and condition RN trials would be greater than that for condition NN trials and for an condition LN and condition NL trials would be smaller than that for condition NN trials.

Results are shown in Figure 5.5. A factorial ANOVA (image $\times$ condition) showed a significant effect of condition on the values of $(F_{h2} - F_{h1})$ ($F(4, 20) = 6.19, p = 0.002$). However, further tests did not reveal the values of $(F_{h2} - F_{h1})$ to be larger for condition NL and condition RN trials than condition NN trials (two-sample one-tailed $t$-test, $p = 0.184$) or the values of $(F_{h2} - F_{h1})$ to be smaller for condition NR and condition LN trials than condition NN trials (two-sample one-tailed $t$-test, $p = 1.000$). There was no significant effect of image on the values of $(F_{h2} - F_{h1})$ ($F(5, 20) = 2.18, p = 0.098$).

As in the previous analysis (Section 5.1.3), KL divergence was also used to test the effect of the de-enhancement. The KL divergence between the horizontal component of fixation locations for the two halves $d_{KL}(F_{h1}, F_{h2})$ was estimated for each condition. If the de-enhancement had affected fixations, the KL divergence between fixation positions in each half of a condition NN trial would be smaller than the values for condition NR, condition RN, condition NL and condition LN trials. Results are shown in Figure 5.6 and demonstrate no effect of the de-enhancement on fixation allocation in this experiment (two-sample one-tailed $t$-test, $p = 0.321$).
CHAPTER 5. STEERING HUMAN FIXATIONS

Figure 5.5: The difference between the mean of horizontal fixation positions under the condition with no de-enhancement and the conditions with a de-enhancement for each image for ramp de-enhancement with change-detection experiment.
Figure 5.6: The KL divergence between horizontal components of fixations recorded during each half of each trial for each image and condition ramp de-enhancement with change-detection experiment.
5.1.5 Ramp de-enhancement with forced-choice target-search task

Image de-enhancement failed to bias fixations away from the region to which it was applied in the last two experiments. In the experiment described in Section 5.1.3 this failure was attributed to the task given to observers or to the fact that observers were viewing the image under the same condition many times. When the task was changed for the experiment described in Section 5.1.4 the de-enhancement still did not have the desired effect. Although the task had been changed, the new task may also have interfered with attempts to steer attention. The observers were searching for an image modification that may have been visible without being viewed directly, so their optimal strategy may not have involved searching for the change with eye movements. To test whether the task in the previous experiment interfered with any influence of the de-enhancement on fixations, the presentation format of Section 5.1.4 was maintained in this next experiment, and a variant on the search task in Section 5.1.3 was introduced.

Observers

Seven observers took part in the experiment (2 male, 5 female, aged 21 to 29 years). All had normal colour vision and normal or corrected-to-normal visual acuity. Observers were not informed of the purpose of the experiment. Two of the observers had also taken part in the experiment described in Chapter 4, one had taken part in the experiment described in Section 5.1.3 and three had taken part in the experiment described in Section 5.1.4.

Procedure

As in the experiment described in Section 5.1.4, observers were presented in each trial with two consecutively displayed images each displayed for 1 s and separated by a 200 ms blank period. The same five trial types as the previous experiment were used,
although in this experiment half the trials were presented were condition NN, and the other half were one of condition NR, condition RN, condition NL or condition LN. The target was present in each trial, superimposed on either the first or second image. After the trial the observers indicated whether the target had appeared in the first or second of the images. Observers performed 1495 – 2795 trials each.

Trials were undertaken in blocks consisting of thirty or thirty-five trials. The calibration routine was improved slightly in this experiment. Calibration data were still recorded at the beginning and end of the experiment and the two sets of calibration data were averaged point-wise and fitted to the target positions by a linear transformation to give least-squares error, but in this experiment calibration data were inspected as they were being recorded. If a valid PoG signal was unavailable for any target, the calibration was repeated. Again, if the calibration error exceeded 0.30 deg the data from the block was not included in analysis and the block was repeated. The mean calibration error for blocks included in analysis was 0.19 deg.

Results and analysis

As before, the method described in Section 3.2.4 was used to extract fixations from the PoG data. Dropped frames, which accounted for 0.81% of the data in this experiment, were removed from the data for each trial before fixations were extracted. Only fixations after the first saccade in each trial half were included in analysis.

As in the previous analysis (Section 5.1.4), the difference between the mean horizontal position of fixations in the second and first half of each trial was calculated. The results are shown in Figure 5.7.

A factorial ANOVA (image × condition) revealed a significant effect of both image and condition on ($\bar{F}_{h2} - \bar{F}_{h1}$) (respectively, $F(4, 20) = 10.16, p < 0.001$ and $F(5, 20) = 10.31, p < 0.001$). The desired effect of de-enhancement was to bias fixations away from the location to which it was applied, which, as has already been mentioned, would
mean that \((\overline{F_{h2}} - \overline{F_{h1}})\) for condition NL and condition RN trials would be larger than condition NN trials, and for an condition LN and condition NR trials would be smaller than NN trials. This was not shown by the data in Figure 5.7 (two-sample one-tailed \(t\)-tests, \(p = 0.192\) and \(p = 0.032\), respectively).

Again, KL divergence was also used to test the effect of the de-enhancement. The KL divergence between the horizontal component of fixation locations for the two halves was estimated for each condition. As before, if the de-enhancement had had an effect on fixations, the KL divergence between fixation position in each half of a condition NN trial would be smaller than the values for condition NR, condition RN, condition NL and condition LN trials. Results are shown in Figure 5.8 and demonstrate no effect of the de-enhancement on fixation allocation in this experiment (two-sample...
Figure 5.8: The KL divergence between horizontal components of fixations recorded during each half of each trial for each image and condition ramp de-enhancement with forced-choice target-search experiment.

one-tailed $t$-test, $p = 0.233$).

### 5.2 Local enhancement

The experiments reported in the previous section showed that despite varying both visual task and presentation format, de-enhancement failed to have a systematic effect on fixations. This could be because the modification was applied to too large an area or because steering attention away from a region using an image modification may simply be unachievable. In the experiments described in this section a different approach was taken. Rather than try to bias fixations away from a large area, an attempt was made to attract fixations towards a smaller area. To this end, the nature of the image modification used in the experiment was altered to local enhancement, which
increased the sharpness of an image region. The weighting function of the image enhancement (described in Section 1.7.6) was changed from a horizontal ramp function to a Gaussian function. Two variants on local enhancement were used: one in which the enhancement appeared abruptly (the dynamic condition) and one in which it was present throughout the trial (the static condition). The logic of using a sudden onset of the enhancement was based on the success shown by McNamara et al. (2008) of a temporal image modification in improving search performance. Four experiments, listed below, were conducted using local enhancement.

1. **Search task with dynamic local image enhancement.** Observers were presented in each trial with an image which appeared for 2 s. When applied, local enhancement appeared abruptly after 1 s. Observers indicated after each trial if they had detected the target.

2. **Detection of dynamic local enhancement.** Observers were presented in each trial with an image which appeared for 2 s. In half the trials local enhancement appeared abruptly after 1 s. Observers indicated after each trial if they had detected any change.

3. **Search task with static local image enhancement.** Observers were presented in each trial with an image which appeared for 1 s. When applied, local enhancement appeared abruptly after 1 s. Observers indicated after each trial if they had detected the target.

4. **Detection of static local enhancement** Observers were presented in each trial with an image which appeared for 1 s. In half the trials local enhancement was applied to the image. Observers indicated after each trial if they had detected any change.
5.2.1 Stimuli

In the experiments in this section, four colour images of natural urban and rural scenes were used. An example of an image with an enhanced region is shown in Figure 5.9 and the remaining three are shown in Figure 5.10. The target used was the same as that used in previous experiments, namely a neutral grey sphere (Munsell N7) subtending approximately 0.3 deg superimposed on the image and matched in luminance to its local surround. The image modification was local enhancement weighted with a Gaussian of standard deviation 0.8 pixels.
5.2.2 Apparatus

The apparatus used in these experiments was the same as that described in Chapter 4 and used in previous experiments.
5.2.3 Search task with dynamic local enhancement

Observers

Twelve observers took part in the experiment (7 male, 5 female, aged 18 to 34 years). All had normal colour vision and normal or corrected-to-normal visual acuity. Observers were not informed of the purpose of the experiment. Four of the observers had taken part in the experiment described in Section 5.1.4 and three had taken part in the experiment described in Section 5.1.5.

Experimental procedure

In each trial an image was presented for 2 s. When present, the enhancement appeared abruptly halfway through the trial. After each trial observers indicated whether they had detected the target. Trials could fall into one of four categories:

1. Target present and enhancement appeared at the same location as the target.
2. Target present and enhancement appeared at a different location from the target.
3. Target absent and enhancement appeared at a random location.
4. Target absent and enhancement did not appear.

The probability of a trial being any of these types was 0.25.

Observers undertook blocks of thirty or thirty-five trials. As the modification was being applied to a smaller area than the modification in Section 5.1, the calibration procedure was altered slightly to ensure that the accuracy of the eye-tracker was maintained. At the beginning of the block, observers undertook a calibration procedure in which they fixated the twenty calibration targets and PoG signals were recorded for each target. These data were fitted to the target positions by a linear transformation, optimized by least-squares. The RMS error was calculated in real time, and if
CHAPTER 5. STEERING HUMAN FIXATIONS

it exceeded 0.30 deg, the procedure was repeated. If an observer failed to achieve a calibration error of less than 0.30 deg the data from that block were not included in analyses. To provide an estimate of any change in an observer’s position, observers undertook another calibration procedure at the end of the block. The PoG signals for each calibration target were transformed to screen co-ordinates by the transformations optimized using the data from the first calibration. If the mean displacement between the two sets of PoG signals exceeded 0.50 deg, the second calibration was repeated. If the mean displacement between the two sets of PoG remained greater than 0.50 deg, data from the block were not included in subsequent analysis. Observers performed 820 – 2080 trials each.

Results and Analysis

The method described in Section 3.2.4 was used to extract fixations from the PoG data. Dropped frames, which accounted for 1.02% of the data in this experiment, were removed from the data for each trial before fixations were extracted. A speed and duration threshold were derived individually for each observer viewing each image.

In order to test if fixations were being attracted by enhancement, data were pooled over all trials in which the enhancement appeared but did not coincide with the target (to avoid a confound caused by observers fixating the target). The positions of each fixation in the second half of each trial (beginning after the enhancement had appeared) relative to the enhancement were pooled. A series of imaginary annuli (thickness 0.5 deg) around the enhancement were defined and the fixation density within each annulus $F_{\text{density}}$ was calculated as the total number of fixations within the annulus divided by its area.

In order to assess the efficacy of the enhancement, a comparison data set was required. A possible way of providing such a comparison set would be to generate randomly located fixations, but since there are well-known biases in the allocation of
fixations such as the central bias and the effect of scene content (Kienzle et al., 2006; Tatler, 2007), this method could be misleading. To produce a comparison data set, fixations from the second half of a different randomly selected trial were substituted for the fixations from the actual trial. If the recorded $F_{\text{density}}$ was higher for the recorded data than those in the comparison data set, it could be concluded that the enhancement was attracting fixations. A plot of $F_{\text{density}}$ versus outer radius of annulus $R_{\text{annulus}}$ is shown in Figure 5.11.

To test the effect of the enhancement on fixations numerically, both fixation density $F_{\text{density}}$ and outer radius of annulus $R_{\text{annulus}}$ were log-transformed and a linear fit made to the data (see Figure 5.11). The gradient of the fit to the recorded fixation density in annuli around the enhancement was steeper than that of the fit to the comparison fixation density in annuli around the enhancement and the intercept with the y-axis was significantly higher for the recorded fixation density data than the comparison data ($z$-tests, $p < 0.001$ for both tests).

In addition to analysis of the eye movement data, observers’ detection performance was analysed using $d'$ (see Section 4.2). Values of $d'$ were calculated for each image independently for trials in which the target and enhancement were together and for trials in which the target and enhancement were apart. Data were pooled over observers. Results are shown in Figure 5.12. As can be seen in the figure, $d'$ was higher when the target and enhancement were together than when they were apart (one-tailed $t$-test, $p = 0.002$).
CHAPTER 5. STEERING HUMAN FIXATIONS

Figure 5.11: The fixation density in annuli around the enhancement for the search task with dynamic local image enhancement experiment (upper figure) and a log-log fit to the same data (lower figure).
Figure 5.12: The detection performance for each image when the enhancement and target were together or apart for the search task with dynamic local image enhancement.
5.2.4 Detection of dynamic local enhancement

In order to test how easily the enhancement could be detected, a control experiment with a reduced number of observers was conducted in which observers were presented with stimuli similar to those used in the previous experiment, but with target always absent, and asked to indicate whether or not they had seen any modification to the image.

Observers

Three observers took part in the experiment (2 male, 1 female, aged 21 to 30 years). All observers had normal or corrected-to-normal vision. Observers were not informed of the purpose of the experiment. One of the observers had also taken part in the experiments described in Sections 5.1.4 and 5.1.5 and two had taken part in the experiment described in Section 5.2.3.

Experimental procedure

In each trial an image was presented for 2 s. In half of the trials, local image enhancement appeared abruptly halfway through the trial. After each trial observers indicated whether they had detected the enhancement. Each image was viewed by each observer in 260 trials.

Results and analysis

The detection performance was calculated for each image. Results are shown in Figure 5.13. Performance for the image enhancement was lower than detection performance for the target (one-tailed $t$-test, $p = 0.019$). This suggests that despite the clear effect of the enhancement upon fixations shown in Figure 5.11, the enhancement itself was not particularly easy to detect.
Figure 5.13: The detection performance for observers attempting to detect the enhancement for each image for the detection of dynamic local enhancement experiment.
5.2.5 Search task with static local image enhancement

Given the success of the attempt to steer fixations in Section 5.2.3, it was of interest to investigate whether fixations were being attracted by the sudden change rather than the contrast modification alone. To this end, a control experiment was conducted in which dynamic local image enhancement was replaced by static local image enhancement.

Observers

Three observers took part in the experiment (1 male, 2 female, aged 18 to 25 years). All had normal colour vision and normal or corrected-to-normal visual acuity. Observers were not informed of the purpose of the experiment. One of the observers had taken part in the experiments described in Sections 5.1.4 and 5.1.5, one had taken part in the experiment described in Section 5.2.4 and all three had taken part in the experiment described in Section 5.2.3

Experimental procedure

In each trial an image was presented for 1 s. When applied, the enhancement was present throughout the trial (with no sudden onset). Observers undertook blocks of thirty or thirty-five trials. As in the experiment described in Section 5.2.3, trials could fall into one of four categories:

1. Target present and enhancement appeared at the same location as the target.
2. Target present and enhancement appeared at a different location from the target.
3. Target absent and enhancement appeared at a random location.
4. Target absent and enhancement did not appear.

The probability of a trial being any of these types was 0.25.
Calibration was undertaken in the same way as in the experiment described in Section 5.2.3. Observers performed 1945 – 2080 trials each.

**Results and analysis**

The method described in Section 3.2.4 was used to extract fixations from the PoG data. Dropped frames, which accounted for 0.35% of the data in this experiment, were removed from the data for each trial before fixations were extracted. Only fixations after the first saccade were included in analysis. As before, in order to test if fixations were being attracted by enhancement, data were pooled over all trials in which the enhancement appeared but did not coincide with the target and the positions of each fixation relative to the enhancement were pooled. A series of imaginary annuli around the enhancement were defined and the fixation density within each annulus was calculated as the total number of fixations within the annulus divided by its area. A similar comparison set to that used in Section 5.2.3 was used, with fixations from a different randomly selected trial being substituted for the fixations from the actual trial. Results are shown in Figure 5.14. As before, both fixation density and distance were log-transformed and a linear fit was made to the data (see Figure 5.14 for the fit). The gradient of the fit to the recorded fixation density in annuli around the enhancement was steeper than that of the fit to the comparison fixation density in annuli around the enhancement and the intercept with the y-axis was significantly higher for the recorded fixation density data than the comparison data (z-tests, $p < 0.001$ for both tests).

Again, observers’ detection performance was analysed using $d'$. Values of $d'$ were calculated for each image independently for trials in which the target and enhancement were together and for trials in which the target and enhancement were apart. Data were pooled over observers. Results are shown in Figure 5.15. As in the data described in Section 5.2.3, $d'$ was higher when the target and enhancement were together than when they were apart (one-tailed $t$-test, $p = 0.001$).
Figure 5.14: The fixation density in annuli around the enhancement for the search task with static local image enhancement experiment (upper figure) and a log-log fit to the same data (lower figure).
Figure 5.15: The detection performance for each image when the enhancement and target were together or apart for the search task with static local image enhancement experiment.
5.2.6 Detection of static local enhancement

In order to test how obvious the static enhancement was, a control experiment was run in which observers were presented with stimuli similar to those used in the previous experiment, but with target always absent, and asked to indicate whether or not they had seen the enhancement.

Observers

Three observers took part in the experiment (2 male, 1 female, aged 21 to 30 years). All had normal colour vision and normal or corrected-to-normal visual acuity. Observers were not informed of the purpose of the experiment. One of the observers had taken part in the experiments described in Sections 5.1.4, 5.1.5 and 5.2.5, two had taken part in the experiment described in Section 5.2.3 and all three had taken part in the experiment described in Section 5.2.4.

Experimental procedure

In each trial an image was presented for 1 s. In half the trials the enhancement was present. When present the enhancement was applied throughout the trial (with no sudden onset). After each trial, observers reported whether they had detected the enhancement. Observers undertook blocks of thirty or thirty-five trials. Each image was viewed in 260 trials.

Results and analysis

The detection performance was calculated for each image. Results are shown in Figure 5.16. Unexpectedly, performance in detecting the enhancement was higher for the static condition than the dynamic condition (one-tailed t-test, $p = 0.038$). The reasons for this were unclear.
Figure 5.16: The detection performance for observers attempting to detect the enhancement for each image for the detection of static local enhancement experiment.
5.3 Discussion

Previously, Einhäuser, Rutishauser and Koch (2008) showed that a search task can override the ability of image modifications to guide fixations and the findings of the experiments in Section 5.1 did not contradict this result. The modification used in those experiments was a de-enhancement which was intended to bias fixations away from an image region, but the desired effect was not achieved. Even when the visual task was altered from searching for a target to explicitly asking observers to perform a more general change-detection task, no systematic effect of the modification was found. This could be attributable to the task (searching for a change in the image was still a search task), but it could also have been caused by the level of the image modification. If observers were consciously aware of de-enhancement every time it was used, they may have developed different strategies in response to the modification. In particular, some may have avoided fixating the de-enhanced region because it was unpleasant to look at; others may have fixated it deliberately as they decided that the target (where present) would be easier to detect on a more featureless background.

Given that attempts to bias attention away from image regions were unsuccessful, the nature of the image modification was altered to local enhancement which subtended a small area and was intended to attract rather than repel fixations. In the first of two conditions using local enhancement, the enhancement appeared suddenly (dynamic local enhancement). The dynamic local enhancement was shown in Section 5.2.3 to achieve the objective of attracting fixations.

Despite improving target detection performance, the enhancement was shown to be generally less detectable than the target. This result shows that a key objective was achieved in this project, that of subtly guiding human fixations. Further to this, the modification attracted attention despite observers being engaged in a search task. It was also shown that when the dynamic local enhancement and target appeared at the
same position, target detection performance was better than when they were separate, presumably because the enhancement attracted fixations.

In the second of two conditions using local enhancement, the dynamic aspect was removed, and the enhancement was present throughout each trial (static local enhancement). This was to test whether it was the sudden image change rather than the image enhancement that attracted fixations. The enhancement continued to draw fixations, and again, when local enhancement and target appeared at the same position, target detection performance was better than when they were separate. The finding that the effect of the static enhancement was similar to that of the dynamic enhancement suggested that it was the enhancement itself rather than the sudden change which was influential in attracting attention.

Although the enhancement was shown to both attract fixations and improve target detection performance, it is possible that this could be due to a form of contextual learning (Einhäuser, Rutishauser and Koch, 2008) rather than because image features had been modified to attract attention by some saliency-type mechanism. The target and enhancement appeared together in half of the target-present trials and this association may have been implicitly learned by observers. Optimum strategy may have involved fixating the enhancement on basis that the target would be at the same location half of the time. A control experiment to test for this would be to repeat the experiments in Sections 5.2.5 with naive observers but include only trials in which the enhancement and target were apart. If the enhancement still attracted fixations then this would rule out contextual learning as a cause for the attraction.

The experiments described here used images in which only edge sharpness had been modified. It is possible that altering other local image features such as chroma or local luminance may also have an effect on fixations. Further work could include an investigation of this hypothesis.
Chapter 6

Summary and Further Work

There were two main objectives in this thesis. The first was to determine which image properties influenced the detectability of a target and the second was to guide human fixations using a subtle image modification. The first objective was addressed in Chapter 4 and the second in Chapter 5. The work in the thesis which was concerned with the human visual system focussed on images of natural scenes (as opposed to synthetic stimuli). Other issues which were addressed included work on information theory and the development of a nonparametric method for fixation classification. Key findings are described in the next section.

6.1 Key results

Factors influencing detection ability in natural scenes

In the investigation of search performance described in this thesis, detection performance was measured at multiple locations on images of natural scenes. The procedure allowed the estimation of local detection performance profiles rather than one global measure for each image. Image features were then ranked in order of the percentage of
variance in local detection performance for which they accounted. The feature which accounted for most variance in detection performance was chroma. It was shown to account for more variance than the higher-level image features such as edge density and Rosenholtz et al.’s (2007) measures of visual clutter.

**Steering visual attention**

An image modification was developed for subtly guiding human fixations. The modification was a local enhancement weighted with a Gaussian function which appeared abruptly halfway through a trial (dynamic local image enhancement). Enhancement succeeded in attracting fixations while observers were engaged in a search task and also improved detection ability when applied to the same location as a target. In an experiment with a reduced number of observers, the dynamic aspect of the modification was removed and the enhancement was present throughout the trial (static local image enhancement). Enhancement continued to attract fixations, which could be argued to show that it was enhancement itself rather than the sudden change that was attracting fixations.

It was not shown to be possible to bias attention from large image regions using an image modification when observers were engaged in either a search task or a change-detection task.

**Non-parametric classification of fixations**

Existing methods for classifying eye movements as fixations or saccades require users to define stability or speed thresholds, a process which is time-consuming and assumes user expertise. Proposed in this thesis is an automatic non-parametric method for classifying eye movements. The method was tested against expert classification and shown to work as well as existing parametric methods, but with the advantage that it did not
require the adjustment of thresholds by hand. The software is to be made publicly available.

**Estimation of KL divergence**

Information theory is a useful statistical tool, but measuring information theoretic quantities is not necessarily straightforward. Methods relying upon the estimation of underlying distributions from data can yield misleading results in some circumstances (such as when data are sparse) so in Chapter 2 a nonparametric method for estimating KL divergence directly from input data (Leonenko et al., 2008) was investigated. When tested using artificial data it was shown to outperform methods which required binning of data or kernel density estimation to estimate underlying distributions.

6.2 Further work

**Testing detection performance using images with a wider variation in high-level properties**

The analyses in Chapter 4 suggested that chroma was a better predictor of local detection performance than high-level image features. As was explained in Section 4.3, this finding mirrored the findings of a study of fixation locations (as opposed to detection performance) by Acık et al. (2009), who concluded that high- and mid-level image features will guide fixations when they are present, otherwise fixations will be guided by low-level features. It could be argued that chroma explained more variance in detection ability than higher-level image features in the experiment described in Chapter 4 simply because there was insufficient variation in high-level image features within each image. An interesting extension of the study would involve repeating the experiment using images which were deliberately selected to have wide variation in image...
features. Other extensions could include altering the properties of the target. For example, a Gabor patch may be more difficult to detect in areas of high edge density than the grey sphere used in the experiments described in this thesis. It is therefore possible that if the experiment were repeated using a target such as a Gabor patch, chroma would become less important to detection than edge density. The colour of the target could also be changed from neutral grey to see if the result remained consistent. An investigation into these factors was not possible within the constraints of this thesis work.

**Could steering fixations be achieved by other low-level image features?**

The experiments described in Chapter 5 showed that fixations could be attracted by the local enhancement of edges in an image. It is possible that this result could extend to other image features such as chroma or luminance.

**Was the effect of the enhancement attributable to learning?**

As was noted in Chapter 5, although the local image enhancement does influence fixations, the mechanism underlying the effect is not immediately clear. The cause may have been a saliency-type mechanism under which the sudden enhancement of the image was visually appealing, and thus attracted fixations. However, the effect could also have been attributed to learning. The target and the enhancement appeared together in half of the target-present trials, and this association may have been learned by observers. The two explanations could be tested by repeating the experiment, but not including any trials where the enhancement and target were together. This was not done in the experiments described here as it was interesting to test the effect of enhancement on detection performance.
More efficient estimation of KL divergence

Leonenko et al.’s (2008) KL divergence estimator gave more stable estimates of the KL divergence between pairs of random vectors than methods based on binning or kernel density estimation. An attempt was made to improve the estimator by producing an offset term which factored out the Gaussian component of the data, but little improvement in convergence was seen. The tests did suggest that it systematically underestimated the true value of KL divergence. This is possibly a consequence of the use of a nearest-neighbours algorithm. It may be possible to reduce the bias of the estimator using a method based on resampling (Efron, 1979; Hall and Martin, 1988).
Bibliography


CIE (2004), *Colorimetry*, 3rd edn, CIE Central Bureau, Vienna, Austria.


Marín-Franch, I. (2009), Information-theoretic analysis of trichromatic images of natural scenes under different phases of daylight, PhD thesis, School of Electrical and Electronic Engineering, University of Manchester, UK.


Wyszecki, G. and Stiles, W. S. (1967), Color Science: Concept and Methods, Quantitative Data and Formulas, John Wiley and Sons, Inc.