# FEATURE SELECTION AND EXTRACTION OF VISUAL SEARCH STRATEGIES WITH EYE TRACKING

**Xiao-Peng Hu**

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
of the
University of London

Department of Computing
Imperial College London
2005

# Abstract

Visual attention is a selective activity of visual search. It stems from the limited ability of our human visual system, and plays an important role in many aspects of visual processing including feature detection, pattern recognition, object identification, and reasoning. The performance of visual search depends heavily on the strategy of deploying visual attention in response to different search tasks. The need to understand visual search strategies has been recognised in both theoretical and practical vision research for many years. Although it has been well established that visual search can be divided into pre-attentive and attentive stages, deterministic models that can predict fixation positions and transitions are not yet available. To date, there has been limited work for identifying the reasons for the choice of fixation points. The method by which observers glean certain subtle features and combine this information with other image features remains poorly understood. The aim of this research is to investigate the theoretical and practical issues involved in visual search strategies for attention selection and knowledge gathering. A new framework for hot spot detection is proposed for capturing the intrinsic visual search behaviour of different observers in image understanding. The method is based on information theory for identifying salient image features used for visual search. We demonstrate how to obtain feature space fixation density functions that are normalised to the image content along the scan paths. This allows the reliable identification of salient image features that can be mapped back to spatial domain for highlighting regions of interest and attention selection. To fully understand visual attention, we also propose an algorithm for extracting the pattern of fixation sequences, which can be crucial for object recognition in order to establish the relationship between visual features and resolve ambiguities. We present a feature selection algorithm based on Bayesian theory for enhancing the understanding of the interaction between top-down and bottom-up processes in human visual perception. We demonstrate the relationship between filter-based feature selection and the performance of ideal inference, in which the expected AUC (the Area under Receiver Operating Characteristic Curve) is employed to describe the intrinsic discriminability of features. A computationally efficient algorithm is designed to reduce the complexity of the feature selection process. With this thesis, we will also propose a Bayesian method for the integration of explicit domain knowledge for scan path analysis, thereby providing a theoretical basis for modelling visual feature selection. The proposed framework is validated with both artificial and real-world

examples, highlighting the robustness and strength of the techniques. The practical value of the method for knowledge gathering for diagnostic decision support in medical imaging is demonstrated with high-resolution Computed Tomography lung image analysis.

*To my family*

# Acknowledgements

This thesis would not have been completed without the support and inspiration of many people.

First and foremost, I would like to thank my supervisor, Professor Guang-Zhong Yang, for his foresight, support, encouragement, and stimulating ideas throughout this research. I would also like to express my gratitude to Professor David Hansell and Dr. Stephen Ellis for their help in organising the experiments involving HRCT images. Special thanks also go to the participants of the eye-tracking experiments.

I am grateful to all members of the Royal Society/Wolfson Medical Image Computing Laboratory at Imperial College London for providing a pleasant working environment and for their constant support and encouragement. I would particularly like to thank Laura Dempere-Marco, whose positive spirit, sense of team-work, and friendship throughout this project is greatly appreciated.

This thesis would not have become possible without the financial support of UK Engineering and Physical Sciences Research Council (EPSRC), and the core funding of the Royal Society/Wolfson Medical Image Computing Laboratory founded by Professor Guang-Zhong Yang. The work reported here is part of the EPSRC Project entitled Visual Tracking for Active Learning – ViTAL (GR/N08810).

Lastly, and most importantly, I wish to thank my parents for their love, support, and constant encouragement. They have been the source of my strength through all the difficult times in my career, especially when I was pursuing my studies abroad. To them I dedicate this thesis.

# Table of Contents

**Abstract** ..... **2**
**Acknowledgements** ..... **5**
**Table of Contents** ..... **6**
**List of Acronyms and Abbreviations** ..... **9**
**List of Figures** ..... **11**
**List of Tables** ..... **16**
**Chapter 1 Introduction** ..... **17**
1.1 Motivation and Objectives ..... 18
1.2 Visual Search for Active Learning and the Scope of the Thesis ..... 20
**Chapter 2 Visual Search: A Computational Modelling Perspective** ..... **26**
2.1 Visual Pathway ..... 27
2.1.1 Basic Optical Pathway ..... 27
2.1.2 Ventral and Dorsal Visual Streams ..... 29
2.1.3 Eye Movement Control ..... 31
2.1.4 Visual Acuity ..... 33
2.2 Experiments for Measuring Visual Search Responses ..... 36
2.2.1 Positron Emission Tomography (PET) ..... 37
2.2.2 Event-Related Potentials (ERP) ..... 37
2.2.3 Functional Magnetic Resonance Imaging (fMRI) ..... 38
2.2.4 Disruption Techniques (Lesion, Damage, Disorder) ..... 38
2.3 Eye Tracking ..... 39
2.4 Stages of Visual Processing ..... 40
2.4.1 Pre-attentive and Attentive Processing ..... 40
2.4.2 The time Course for Visual Processing ..... 41
2.4.3 Attention Selection ..... 41
2.4.4 Space, Feature and Object-Based Selection ..... 43
2.4.5 Attention Deployment and Eye Movement ..... 44
2.5 Information Representation ..... 45
2.5.1 Bottom-Up (Image-Based, Task-Independent) Information ..... 45
2.5.2 Top-Down (Task-Dependent) Information ..... 46
2.5.3 Visual Memory ..... 47
2.5.4 The Binding Problem ..... 47
2.5.5 Object Recognition ..... 48
2.6 Computational Modelling and Simulation ..... 49
2.6.1 Feature Extraction and Feature Maps ..... 50
2.6.2 Feature Combination and the Saliency Map ..... 51
2.6.3 Attention-selection Control ..... 52
2.6.4 Top-Down Activity ..... 55
2.6.5 System Validation ..... 56
2.6.6 Signal Detection Theory ..... 56

2.7 Summary ..... 57
**Chapter 3 Scan Path Analysis and Feature Extraction ..... 58**
3.1 Eye Tracking and Fixation Analysis ..... 59
3.1.1 Eye Tracking ..... 59
3.1.2 Fixation Analysis ..... 61
3.2 Image Feature Extraction ..... 65
3.2.1 Texture and Texture-Feature Extraction Techniques ..... 65
3.2.2 The Segmentation of Anatomical Structures ..... 68
3.2.3 Feature Extraction ..... 71
3.3 Conclusions ..... 77
**Chapter 4 Framework of Hot Spot Detection ..... 78**
4.1 Material and Methods ..... 80
4.1.1 Representation of Visual Search Scan Path ..... 80
4.1.2 The Fixation Density Function ..... 81
4.1.3 Hot Spot Analysis ..... 83
4.2 Experiment Setup and Results ..... 88
4.2.1 Experiment Setup ..... 88
4.2.2 Two-Colour Conjunction Search ..... 88
4.2.3 Feature Selection and Hot Spot Detection for CT Images of the Lungs ..... 91
4.3 Discussion and Summary ..... 97
**Chapter 5 Feature Space Analysis of Fixation Sequences .... 100**
5.1 Material and Methods ..... 101
5.1.1 Feature Space Representation of Fixation Sequences ..... 101
5.1.2 The Transition Density Matrix ..... 102
5.1.3 Feature Space Analysis of Fixation Sequences ..... 104
5.2 Experimental Results ..... 107
5.2.1 Two-Colour Conjunction Search ..... 107
5.2.2 Conjunction Search ..... 109
5.2.3 Fixation Sequence Analysis for CT Images of the Lungs ..... 114
5.3 Discussion and Summary ..... 116
**Chapter 6 Bayesian Framework for Feature Selection ..... 118**
6.1 Feature Relevance and Feature Selection ..... 119
6.2 The Theoretical Framework ..... 122
6.2.1 Irrelevance Definition ..... 122
6.2.2 Feature Selection and the Performance of Decision-Making ..... 123
6.2.3 Feature Selection Based on ROC Analysis ..... 130
6.3 Experiment Validation ..... 133
6.4 Conclusions ..... 137
**Chapter 7 Bayesian Modelling for Visual Feature Selection and Attention Deployment during Visual Search ..... 139**
7.1 Bayesian Modelling of Visual Attention ..... 141
7.1.1 Bayesian Modelling ..... 141
7.1.2 Visual Feature Selection ..... 144
7.1.3 Feature Extraction and Probability Density Estimation ..... 145
7.1.4 Scan Path Analysis for Modelling Attention Selection ..... 149
7.1.5 Spatial Attention Selection ..... 150
7.2 Experimental Results ..... 151

7.2.1 Two-Colour Conjunction Search ..... 151
7.2.2 Real-World Applications ..... 153
7.3 Discussion and Summary ..... 166
**Chapter 8 Conclusions and Future Work ..... 168**
8.1 Summary of the Thesis ..... 168
8.2 Future Work ..... 172
**Appendix A Proofs for Theorems 6.1 and 6.2 ..... 176**
**References ..... 183**

# List of Acronyms and Abbreviations

| | |
|---|---|
| AIT | Anterior Infero-Temporal cortex |
| aROI | algorithmically detected Region of Interest |
| ASL | Applied Science Laboratories |
| AUC | Area Under Receiver-Operating-Characteristic Curve |
| BOLD | Blood-Oxygen-Level-Dependent |
| CR | Corneal Reflection |
| CT | Computed Tomography |
| DOG | Difference-Of-Gaussian |
| EBA | Extrastriate Body Area |
| EM | Expectation-Maximisation |
| EOG | Electro-Oculography |
| EPSRC | Engineering and Physical Sciences Research Council (UK) |
| ERP | Event-Related Potentials |
| ERS | Edge Radius Symmetry |
| FEF | Frontal Eye Field |
| FFA | Fusiform Face Area |
| FIT | Feature Integration Theory |
| fMRI | functional Magnetic Resonance Imaging |
| glu | grey-level uniformity |
| HRCT | High Resolution Computed Tomography |
| hROI | human identified Region of Interest |
| IT | Infero-Temporal cortex |
| KL | Kullback-Leibler |
| LGN | Lateral Geniculate Nucleus |
| LIP | Lateral Intra-Parietal area |
| LLE | Local Linear Embedding |
| LO/LOC | Lateral Occipital Complex |
| LOI | Locally Orderless Image |
| lpe | long primitive emphasis |
| MDS | Multi-Dimensional Scaling |
| ML | Maximum Likelihood |

| | |
|---|---|
| NB | Naïve Bayes |
| PCA | Principal Component Analysis |
| PET | Positron Emission Tomography |
| PFC | Pre-Frontal Cortex |
| PIT | Posterior Infero-Temporal cortex |
| plu | primitive length uniformity |
| POG | Photo-Oculography |
| PPA | Parahippocampal Place Area |
| PSF | Point Spread Function |
| RLL | Right Lower Lobe |
| RML | Right Middle Lobe |
| ROC | Receiver Operating Characteristic |
| RT | Response Time |
| RUL | Right Upper Lobe |
| SC | Superior Colliculus |
| SDT | Signal Detection Theory |
| spe | short primitive emphasis |
| SPL | Superior Parietal Lobule |
| V1 | Primary Visual Cortex |
| V4 | Fourth Visual Area |
| V5/MT | Middle Temporal area |
| VIP | Ventral Intra-Parietal area |
| ViTAL | Visual Tracking for Active Learning |
| VOG | Video-Oculography |

# List of Figures

**Figure 1.1** The general framework of Visual Tracking for Active Learning (ViTAL)..... 21
**Figure 1.2** A framework for visual search strategy analysis. Explicit domain knowledge for this task was utilised to define the dimension of the feature space..... 21
**Figure 2.1** A cross section of the retina (*Adapted from D.H. Hubel's Eye, Brain, and Vision. W.H. Freeman, New York, 1995*)..... 28
**Figure 2.2** A schematic illustration of the basic optical pathway (*Adapted from D.H. Hubel's Eye, Brain, and Vision. W.H. Freeman, New York, 1995*)..... 29
**Figure 2.3** A schematic illustration of the dorsal and ventral visual streams..... 30
**Figure 2.4** A diagrammatic overview of the extrinsic muscles of the eye. The *medial* and *lateral recti* control sideway movements; the *superior* and *inferior recti* control up/down movements; the *superior* and *inferior obliques* control twist movements. 1: superior rectus; 2: levator palbebrae superioris; 3: lateral rectus; 4: medial rectus; 5: superior oblique; 6: reflected tendon of the superior oblique; 7: annulus of Zinn; 8: inferior rectus; 9: inferior oblique (*Adapted from H. Davson's Physiology of the Eye. 4th ed, Academic Press, New York, 1980*)..... 32
**Figure 2.5** A distribution of cones (dotted lines) and rods (solid lines) on the human retina (*Adapted from R.N. Haber and M. Hershenson's The Psychology of Visual Perception. Holt, Rinehart, and Winston, Inc., New York, 1973*)..... 34
**Figure 2.6** The effect of eccentricity on visual acuity (*Adapted from G. Westheimer's Visual Acuity. Chapter 17. In: Moses, R. A. and Hart, W. M. (ed) Adler's Physiology of the eye, Clinical Application. St. Louis: The C.V. Mosby Company, 1987*)..... 35
**Figure 2.7** Patterns to illustrate the serial/parallel dichotomy in visual search. It is easier to find 'X' in the right pattern than 'L' in the left pattern..... 36
**Figure 2.8** (a) The basic principle of how the four Purkinje images are formed. (b) The relative positions of the pupil and the first Purkinje reflection during visual search as seen from the camera when the observer fixates to the centre and the four corners of the screen. (*Adapted from Visual search: psychophysical models and practical applications, by G.Z. Yang, L. Dempere-Marco, X.P. Hu and A. Rowe, Image and Vision Computing, vol. 20, pp. 291-305, 2002*)..... 40
**Figure 2.9** A typical scan-path recorded by an eye-tracker. The search task is to identify a pair of father and son. The size of circles represents the dwell time of the fixations whereas the lines between fixations indicate the saccadic eye movements..... 44
**Figure 2.10** A feed-forward model for object recognition..... 49
**Figure 2.11** A typical computational model to demonstrate its basic components..... 50
**Figure 2.12** A schematic illustration of the relationship between feature maps and the saliency map. Visual features are independently registered in feature maps. Each unit in the saliency map is arranged in one to one correspondence with the units of the retinal image. The value of each unit in saliency map reflects the spatial competition of visual features, registered in each feature map..... 51

**Figure 2.13** The rigid-spotlight attentional model ..... 53
**Figure 2.14** Activity of four attentional units. Unit 1 (black line) has the biggest external input. Unit 3 and unit 4 (two coinciding green lines) have the smallest external input. Unit 2 (red line) has a medium external input. ..... 53
**Figure 3.1** The ASL Model 504 remote eye-tracking system. (a) The infrared camera used to capture the eye image. The white circle on (b) indicates the pupil. The white and black crosses indicate the centre of the pupil and the corneal reflection respectively. The direction of eye gaze is calculated from the relative positions of these reflections ..... 60
**Figure 3.2** A schematic diagram of the DICOM image-viewing platform ..... 60
**Figure 3.3** Flow chart of the fixation calculation algorithm. *Th*1, *Th*2 and *Th*3 are three boundary thresholds, which are 0.5, 1.0 and 1.5 degrees respectively. The set $S_{fix}$ stores the gaze points of the current fixation. The set $S^0_{outside}$ is obtained by using *Th*3 to filtering out noise points in $S_{outside}$. The function *Angle*(*P*) calculates the visual angle of the gaze point *P*; *Mean*(*S*) calculates the mean value of the visual angles of the gaze points in the gaze point set *S*; *StdDev*(*S*) calculates the standard deviation of the visual angles of the gaze points in *S*; *Number*(*S*) calculates the number of the gaze points in *S* ..... 63
**Figure 3.4** Results of the fixation algorithm shown in Figure 3.3. The eye movements were recorded when an observer was asked to search for a waving hand in picture (a). (b) and (c) illustrate the results of the fixation calculation by setting the three boundary thresholds to (0.5, 1.0, 1.5) and (0.75, 1.25, 1.75) degrees, respectively. The red dots indicate the recorded eye gaze positions; the size of circles represents the dwell time of each fixation and the lines between fixations indicate the saccadic movements. It can be seen that some large fixations in (c) are separated into several adjacent small fixations, indicating the effect of the boundary thresholds ..... 64
**Figure 3.5** (a) A schematic illustration of the anatomy of the lungs (including the trachea). (b) The bronchial tree (*Adapted from Corel Gallery™ Magic 65,000*) ..... 69
**Figure 3.6** Segmentation of the lung parenchyma and blood vessels. (a) original HRCT slice (b) the extracted parenchyma (c) the blood vessels marked with red colour ..... 70
**Figure 3.7** A CT image and its extracted texture features by using the sixteen feature extractors summarised in Table 3.1 ..... 73
**Figure 3.8** The 25 'first order' filters used in computing the feature maps, numbered from 0 to 24. ..... 75
**Figure 3.9** 20 selected 'texture of texture' feature maps of the CT image. The number pairs under each map indicate how the maps are generated. For instance, '23-15' indicates that the map is generated by applying filter 23 and filter 15 (shown in Figure 3.8) sequentially ..... 76
**Figure 4.1** The overall framework of the proposed method for feature and attention selection based on visual search scan path analysis. Explicit domain knowledge for this task was utilised to define texture dimension for the feature space. ..... 84
**Figure 4.2** A two-colour conjunction search experiment used for illustrating the key steps involved in the proposed framework. Each small square in the image contains two different colours. The observers were asked to identify the square that contains both green and yellow. Data recording was stopped as soon as the observers had successfully identified the target. The size of the circles represents the dwell time of each fixation whereas the lines between fixations indicate the saccadic eye movements. (a) and (b) demonstrate the different search patterns adopted by one observer. The recorded scan paths can be used to extract the underlying strategies used. ..... 89

**Figure 4.3** Feature space histograms of $T(f_k)$, $T_0(f_k)$ and $\Gamma(f_k)$ for the scan paths used in Figures 4.2(a) and (b). The fixation density $\Gamma(f_k)$ in (b) has a larger value on green, which indicates that green is where the attention has been focussed upon while searching for the target. These figures also indicate that attention selection in feature space cannot be detected by using total dwell time $T(f_k)$ alone. ..... 90
**Figure 4.4** The temporal changes of fixation density function $\Gamma(f_k)$ for the scan paths used in Figures 4.2(a) and (b), illustrating the changes/convergence of visual search strategy. ..... 91
**Figure 4.5** Four HRCT lung images of a patient with centrilobular emphysema and the corresponding visual search scan paths recorded from two experienced radiologists. The size of the circles represents the dwell time of each fixation and the lines between fixations indicate the saccadic eye movements. ..... 92
**Figure 4.6** Feature space representation of $T(\cdot)$, $T_0(\cdot)$ and $\Gamma(\cdot)$ corresponding to scan path 1 of radiologist A. The peak of $\Gamma(\cdot)$, with feature values ranging from 6.0 to 15.0, represents hot features and can be back-projected onto the original HRCT lung images, indicating regions of interest where visual attention should be directed towards. The highlighted areas of the four HRCT images are obtained by back-projection of the feature Co-Maximum with value from 6.0 to 15.0. ..... 94
**Figure 4.7** The fixation density function $\Gamma(\cdot)$ of scan path 1 recorded from radiologist B in the feature space formed by texture features of absolute deviation and skewness. The two highlighted regions can subsequently be back-projected to the original HRCT lung images. ..... 95
**Figure 4.8** Back-projections for image 1 from the detected hot spots in feature space with scan paths recorded from the two radiologists for images 1-4 respectively. ..... 96
**Figure 4.9** Combined analysis of regions of interest from the two observers by following the Bernoulli trial, where brighter grey shades within the lung parenchyma in (a) and (b) signify higher estimated probabilities and variance, respectively. Those pixels highlighted in the images of (c) have an estimated probability greater than 50%. ..... 96
**Figure 5.1** The integrated transition patterns of the scan paths in Figures 4.2(a) and (b). The highlighted values are the largest values outwards each state. The fact that all the values towards green are highlighted in (b) indicates that visual attention is guided to the green colour while searching for the target in Figure 4.2(b). ..... 109
**Figure 5.2** The image designed for examining how comparisons affect scan paths. The polygons can have different sizes (small or large), different colours (black, red, yellow or green) and different shape (3, 4, 5 or 6 edges). The targets are a pair of father and son polygons that have the same colour and shape but are different in size. ..... 110
**Figure 5.3** Two scan paths acquired from two observers. The size of the circles represents the dwell time of each fixation, and the lines between fixations indicate the saccadic eye movements. ..... 111
**Figure 5.4** Four HRCT lung images of a patient with centrilobular emphysema and the corresponding visual search scan paths recorded from two experienced radiologists. The two highlighted regions in (c) are obtained by the back-projection of hot feature spots. ..... 115
**Figure 6.1** A typical model for feature selection. ..... 120
**Figure 6.2** The search space for selecting feature subset from three input features. Each state in the space represents a candidate feature subset. For instance, state 101 indicates that the second feature is not included. ..... 120

**Figure 6.3** Conditional probabilities and their ROC curve. Two conditional probability functions are plotted in (a). In the ROC space shown in (b), hit rate change s from 0% to 100% when false-alarm rate varies from 0% to 100%..... 124
**Figure 6.4** A graphical illustration of the discriminability of features. Each region bordered by an inner curve and the outer circle represents the discriminability of a feature. There can be overlaps between features. The overall discriminability is represented by the area of the region bordered by the outer circle..... 128
**Figure 7.1** A schematic diagram of the Bayesian modelling of visual attention selection. The input image is decomposed in a parallel fashion into visual features, among which relevant features are selected. Saliency map and top-down control are then computed in the selected feature space. Saliency map drives visual attention to regions with rare features, whereas top-down activation directs visual attention towards regions sharing similar features of the target. The bottom-up saliency is modulated with the top-down bias into a coherent activation map, based on which the input image is sequentially attended in an order of decreasing activation..... 143
**Figure 7.2** A two-colour conjunction search experiment used for illustrating the key steps involved in the proposed framework. Each small square in the image contains two different colours. The observers were asked to identify the square that contains both green and yellow. Data recording was stopped as soon as the observers had successfully identified the target. The size of the circles represents the dwell time of each fixation whereas the lines between fixations indicate the saccadic eye movements. The primary colour used by the observers for locating the target is green in this experiment..... 152
**Figure 7.3** The change of the AUC in the backward feature selection. The AUC value remains unchanged after red and black are eliminated. When yellow is removed, the AUC value decreases slightly from 0.997 to 0.949. 'bryg' denotes that the corresponding AUC is achieved by using the black, red, yellow and green features..... 152
**Figure 7.4** A Simulated scan path. Green is selected by the feature selection algorithm and is used to direct visual attention..... 153
**Figure 7.5** The HRCT lung image of a patient with centrilobular emphysema and the corresponding visual search scan path recorded from an experienced radiologist. The size of the circles represents the dwell time of each fixation and the lines between fixations indicate the saccadic eye movements..... 154
**Figure 7.6** The top-down, bottom-up and activation maps of the CT image shown in Figure 7.5. The number of selected features is 20, 30 and 40 respectively. Brighter grey shades within the lung parenchyma in (a), (b) and (c) signify higher values. Those pixels highlighted in the images of (d) have an activation value greater than 1.0..... 155
**Figure 7.7** The simulated scan paths by using 20, 30 and 40 selected features of the CT image shown in Figure 7.5. The circles represent fixations and the lines between fixations indicate the saccadic eye movements..... 156
**Figure 7.8** Another CT image was obtained from the same patient of Figure 7.5. It is different from the CT image shown in Figure 7.5 in the shape of the parenchyma as well as the spatial distribution of the 'black lung' regions..... 156
**Figure 7.9** The top-down, bottom-up and activation maps of the CT image shown in Figure 7.8. The number of selected features is 20, 30 and 40 respectively. Brighter grey shades within the lung parenchyma in (a), (b) and (c) signify higher values. Those pixels highlighted in the images of (d) have an activation value greater than 1.0..... 157

**Figure 7.10** The simulated scan paths by using 20, 30 and 40 selected features of the CT image shown in Figure 7.8. The circles represent fixations and the lines between fixations indicate the saccadic eye movements. ..... 158
**Figure 7.11** The CT image shown in Figure 7.8 is identical to the above CT image. The superimposed scan path was recorded from the same radiologist of the scan path shown in Figure 7.5. The number of selected features is 20, 30 and 40. Those pixels highlighted in the images of (b) have an activation value greater than 1.0. ..... 159
**Figure 7.12** A satellite image of a part of Beijing city and two superimposed scan paths. Scan path (a) was obtained when an observer was asked to search for the details of the city, while scan path (b) was obtained when the observer was asked to view the mountain appearance. ..... 161
**Figure 7.13** Three satellite images of three different parts of Beijing city. ..... 161
**Figure 7.14** The activation maps and simulated scan paths corresponding to the scan path shown in Figure 7.12(a) for the images shown in Figure 7.13 by using 20 selected features. Brighter grey shades in the activation maps signify higher values. Those pixels highlighted in the images of the second panel have an activation value greater than 1.0 ..... 162
**Figure 7.15** The activation maps and simulated scan paths corresponding to the scan path shown in Figure 7.12(b) for the three images shown in Figure 7.13 by using 20 selected features. Brighter grey shades in the activation maps signify higher values. Those pixels highlighted in the images of the second panel have an activation value greater than 1.0. ..... 163
**Figure 7.16** The superimposed scan paths were recorded from the same observer of the scan path shown in Figure 7.12 when searching for the details of the city. The number of selected features is 20. Brighter grey shades in the activation maps signify higher values. Those pixels highlighted in the images of the bottom panel have an activation value greater than 1.0. ..... 164
**Figure 7.17** The superimposed scan paths were recorded from the same observer of the scan path shown in Figure 7.12 when searching for the details of the mountains. The number of selected features is 20. Brighter grey shades in the activation maps signify higher values. Those pixels highlighted in the images of the bottom panel have an activation value greater than 1.0. ..... 165

# List of Tables

**Table 2.1** A representative time course for rapid visual processing in the brain. ..... 41
**Table 3.1** A summary of the 16 texture features used for analysing the HRCT images of the lung ..... 71
**Table 4.1** The probabilities of the colours being hot, variance and 90% confidence interval of the two-colour conjunction search experiment for the twelve subjects studied ..... 91
**Table 4.2** *KL* Divergences between $T(\bullet)$ and $T_0(\bullet)$ in the feature space for the scan paths made by two observers when assessing the HRCT images. The highlighted values are the largest values of the columns ..... 93
**Table 4.3** The selected texture features for scan path analysis in 1D feature space and their corresponding *KL* Divergence. Both scan paths 1 and 4 of observer B have rather small *KL* values (highlighted), which indicates that additional features should be considered in determining hot spots during visual search. ..... 93
**Table 6.1** Summary of the seven datasets used for validation. Corral, LED24 and Monk2-Local are artificial datasets, and others are real-world datasets. The baseline accuracy for an artificial dataset is the frequency of the majority class on the test dataset, and the baseline accuracy for a real-world dataset is the frequency of the majority class on the whole dataset. ..... 133
**Table 6.2** The CPU time (seconds) for feature selection on a Pentium III (CPU cache: 512KB, CPU frequency: 450MHz, Memory: 384MB) computer under SuSE Linux 7.2 operating system ..... 133
**Table 6.3** The results of feature selection on artificial datasets. The accuracy is the ratio of the number of relevant features selected against the total number of selected features. The expected AUC in the Table is the expected AUC of selected features ..... 134
**Table 6.4** A comparison of NB and C4.5 with and without feature selection, where the class probabilities are estimated from data. Accuracy is expressed as the percentage error rate for each classifier. ..... 135
**Table 6.5** A comparison of NB and C4.5 with and without feature selection, where the class probabilities are set to be equal. Accuracy is expressed as the percentage error rate for each classifier. ..... 136
**Table 7.1** Results of comparison between the activation maps shown in Figures 7.9 and 7.11. ..... 160
**Table 7.2** Results of comparison between the activation maps shown in Figures 7.14 and 7.16. ..... 166
**Table 7.3** Results of comparison between the activation maps shown in Figures 7.15 and 7.17. ..... 166

# Chapter 1

# Introduction

Visual search often involves the task of finding targets among distractors. Generally, the process of visual search can be divided into two stages: pre-attentive search and attentive search [1]. During the pre-attentive stage, basic visual features (*e.g.* stimuli such as colour, orientation, motion and size) over the entire visual field are independently registered into the visual memory in a parallel fashion. When information about the relationship between features is required, the attentive search, which is often a slower and serial scan over different parts of the visual field, is necessary [2]. Visual attention, which stems from the limited ability of our human visual system, plays an important role in many aspects of visual processing, including feature detection, pattern recognition, object identification, and memory.

The acuity of our eyes is not uniformly distributed. Instead, the best visual acuity is only within a visual angle of one to two degrees. As a result, it is necessary for the eyes to scan the scene actively with a series of fixations and saccades. A fixation is an ocular dwell over a location for a certain period of time, whereas saccades are fast, ballistic voluntary eye movements from one fixation to another for building up a detailed representation of the scene. Overt eye movement may be different from visual attention selection, but they are very closely related [1]. Since information is gathered mainly during fixations rather than saccades, it is possible to obtain a glimpse of the cognitive process that governs visual search tasks by analysing where eye gaze has been directed.

## 1.1 Motivation and Objectives

In recent years, there have been rapid advances in medical imaging techniques with a significant increase in both the volume and the quality of data generated. In parallel with these advances, the expectation of the reliability of medical diagnosis has also changed. Decision support systems are playing an increasingly important role in clinical diagnosis, training and quality assurance. Hitherto, clinical diagnosis for medical images acquired from different modalities has largely been a subjective process, which can result in considerable inter- and intra-observer variability [3]. Although the value of decision support systems in medical imaging has been appreciated for many years, the existing techniques are mainly based on traditional artificial intelligence approaches [4, 5]. Explicit domain knowledge, which in this case is represented as low-level visual features, is integrated into the system along with *ad hoc* feature extractors and decision rules. Such an approach has major drawbacks in that human visual pattern recognition is different from general reasoning. Explicit domain knowledge representation often overlooks those factors that are subconsciously applied during visual recognition. Domain knowledge acquisition is a time-consuming task, and the *ad hoc* nature of grouping low-level visual features means that there is no coherent approach in the development of decision support systems. These drawbacks call for the development of a new framework for a natural and systematic way of gathering knowledge from domain experts.

Studies on visual tracking have revealed that scan paths generated from eye movements usually have repetitive and idiosyncratic patterns [6, 7]. In radiology, for instance, when radiologists are presented with an image, they do not scan it systematically but rather concentrate their foveal vision on a number of regions of interest. Experienced radiologists seem to investigate fewer fixation points than unskilled observers [8, 9]. The nature of this discrepancy is context dependent and governed by the experience, knowledge and expectation of the observers. Eye movements between fixations are purposeful and refined by training. It is therefore possible to investigate the underlying visual search strategies employed by medical experts through eye tracking.

From an optimisation perspective, a strategy is a guiding principle for achieving the best performance (or optimising behaviour) with respect to the external condition given limited resources [10]. The strategy may not be directly observable but could be inferred from repeated indirect observations. In the analysis of the scan path of visual search, the strategy of attention deployment refers to attention selection and sequences. The selection

characteristics of the strategy are directly related to the basic question of 'what are we looking at?' A growing body of studies suggests that attention selection can operate on spatial locations, visual features or objects. To achieve optimal performance, we tend to select salient objects within a scene. The saliency of an item is governed by its bottom-up aspects, such as the spatial arrangement of visual stimuli and their uniqueness, and its top-down aspects, which are associated with the knowledge, interests and expectation of the observer. On the other hand, the sequential aspects of the visual search strategy are related to the question of 'where should we direct our visual attention?' when several salient items are available. This sequence can be crucial in object recognition for establishing the relationship between visual features/objects.

The need to discover visual search strategies has been recognised in both theoretical and practical vision research for many years [4]. Although it is well established that visual search can be divided into pre-attentive and attentive search stages [1], deterministic models to predict fixation positions and transitions are not yet available. To date, limited work has been conducted in identifying the reasons for the choice of fixation points, or on making use of the information gained from fixation points and sequences. The method by which an experienced radiologist gleans and combines certain subtle features remains poorly understood.

The aim of this research is to investigate the theoretical and practical issues involved in discovering visual search strategies for attention selection in image understanding. By making direct use of the way that medical experts perform visual examinations, we aim to provide a general framework for knowledge gathering based on visual search. This is useful not only for building a visual diagnostic decision support system but also for the training and continuous assessment of radiologists. This framework will facilitate the systematic definition of diagnostic features and the extraction of relevant knowledge from a visual domain.

In addition to its applications in medical decision support systems, the modelling of visual search strategies is also significant for a number of other computer vision applications, such as those listed below:

- *Image and video compression*: regions of interest are transmitted or stored with different visual acuity depending on visual attention [11].

- *Active machine vision*: the active nature of the human vision systems is simulated to facilitate the design of machine vision systems [12, 13].
- *Image retrieval, object recognition and machine inspection*: the use of feature transitions between regions of interest for defining semantic contents of the scene [14].

## 1.2 Visual Search for Active Learning and the Scope of the Thesis

The work reported in this thesis is a part of the ViTAL (Visual Tracking for Active Learning) framework being developed at Imperial College. The goal of ViTAL is to provide a general framework for knowledge gathering for diagnostic decision support systems [15]. We aim to make use of the spatio-temporal characteristics of human visual search, together with the intrinsic visual features, as the basis for domain knowledge representation. The general methodology for incorporating fixation points and subliminal cross-checking of the intra- and inter-features for knowledge gathering is not dependent on a single disease or image modality, and should be applicable to a wide range of decision support applications that involve digital imaging.

The basic structure of ViTAL is illustrated in Figure 1.1. An eye tracker is used for collecting spatio-temporal information of eye movements during normal, uninterrupted radiological examination sessions by experienced observers. This information is then used for spatio-temporal analysis of visual search patterns based on fixation-triggered local feature analysis. With this framework, the dynamic information about the dwell time at fixation points and the temporal order with which the inter- and the intra- image features are compared are used to form a decision support network. The local feature analysis engine provides a feedback to the observer for interactive validation of the importance of different low-level image features during visual search. With this framework, explicit domain knowledge and non-imaging related physical measurements are typically incorporated as constraints, *e.g.* lung function test data in the case of diagnosing diffuse lung disease.

The essential part of the ViTAL framework is knowledge gathering, which is also the emphasis of the work presented in this thesis. Based on the general structure of ViTAL,

we propose a framework as shown in Figure 1.2 to discover visual search strategies for attention selection and the patterns of fixation sequences in the feature space.

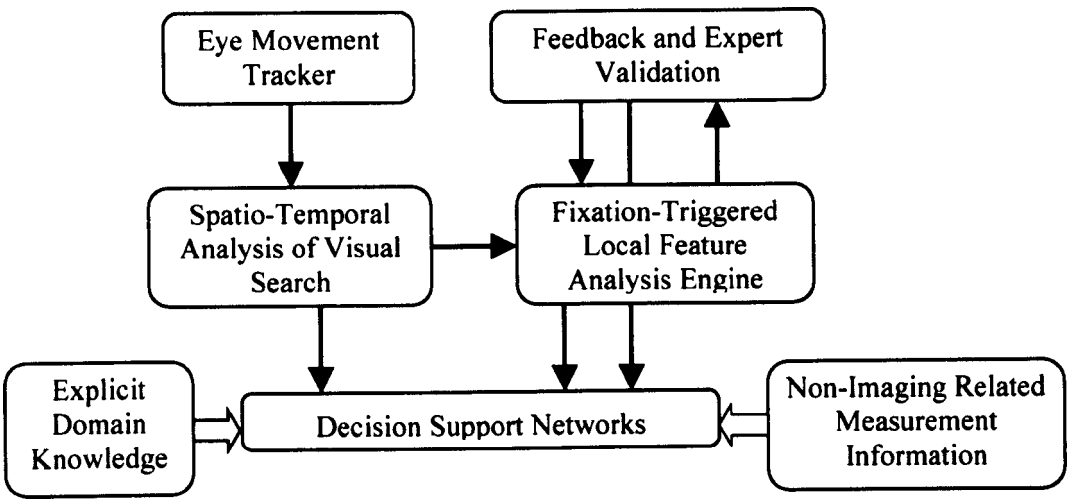



**Figure 1.1** The general framework of Visual Tracking for Active Learning (ViTAL).

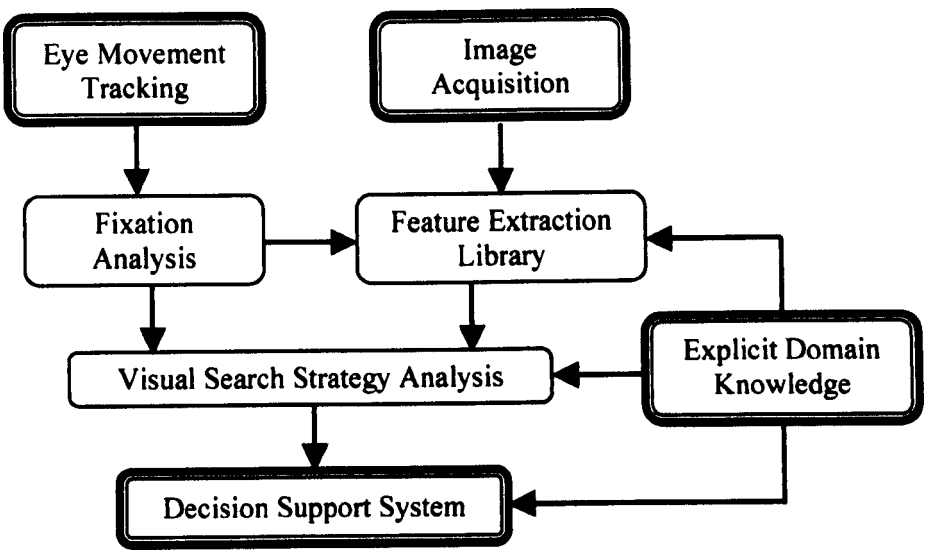



**Figure 1.2** A framework for visual search strategy analysis. Explicit domain knowledge for this task was utilised to define the dimension of the feature space.

In the above framework, fixations and saccadic eye movements are analysed through spatio-temporal clustering in both spatial and feature spaces. In order to extract regions of interest in the feature space, the information of fixation points is subsequently fed into the feature extraction library, which is generic and domain independent. Visual search strategies are then analysed by integrating image contents (characterised by visual

features) and fixation patterns. The design of the above framework raises the following essential questions:

- How can we extract and select suitable visual features to construct a feature space in which visual search is performed?
- How can we define a suitable representation of the information related to visual attention selection in the constructed feature space?
- How can we define a meaningful representation of the fixation sequences in the constructed feature space?
- How can we integrate search patterns obtained from different observers?
- How can we use the acquired information to highlight regions of interest and provide a simulated scan path for guidance in decision support?

In order to answer the above questions, this thesis is organised as follows. In Chapter 2, an overview of the basic principles of visual search studies is provided. Particular emphasis is placed on the computational modelling of visual attention. Issues related to the visual pathway, including ventral and dorsal streams, eye movement control and visual acuity, are introduced. The basic concept of human visual processing and how it can be applied to the construction of a biologically plausible computational model of visual attention are also discussed.

The methods for scan path analysis and feature extraction are described in Chapter 3. After the introduction of a DICOM image-viewing platform for recording eye movements, a fixation calculation algorithm based on spatio-temporal analysis of recorded eye movements is discussed. Both *ad-hoc* and generic texture features are extracted to form an image feature library. In this chapter, we also describe the basic steps involved in the segmentation of basic anatomical structures of the lung, including the chest wall, lung parenchyma and macroscopic pulmonary vessels, together with a brief survey of texture analysis approaches.

Chapter 4 presents a framework of hot spot detection, which is designed to tackle attention selection during visual search. In this framework, scan paths are first projected into feature space to detect 'hot' feature regions, which can be projected back after the projection bias (which is independent of visual search strategies) is eliminated by using feature space fixation density functions. A simple forward feature selection based on the Kullback-Leibler divergence is also discussed. We demonstrate how the common patterns

embedded in the scan paths acquired from different observers can be detected through an integration procedure, which can be approximated by a Bernoulli test. The potential value of the method is assessed by both artificial and medical imaging experiments.

A method for identifying the patterns of fixation sequences *via* feature space analysis is proposed in Chapter 5. With this approach, the states for fixation sequences analysis are derived from the results from hot spot detection. A conjunction object search experiment is designed to demonstrate how the proposed method works. The results illustrate the difficulties of feature comparison in visual search, highlighting the importance of training and knowledge gathering in medical imaging.

The theoretical and practical issues about feature selection are further discussed in Chapter 6. After a brief introduction of feature selection methods, we demonstrate the relationship between *ROC* analysis and the dimensionality of a feature space. A filter-based feature selection method is proposed, and we provide an approximation algorithm that is designed to reduce the computational cost. Validation results by using both artificial and real-world data sets indicate that the algorithm is able to remove irrelevant and redundant features and, in many cases, the classification performance can be improved.

Chapter 7 presents a computational framework for modelling attention deployment based on Bayesian theory and a biologically plausible architecture. With this framework, feature selection is regarded as a consequence of interaction between top-down and bottom-up process. The algorithm developed in Chapter 6 is applied for optimal visual feature selection. The possibility of incorporating human visual search behaviours (detected by eye movement analysis) into scan path simulation is explored. For system validation, Spearman correlation is applied to compare saliency maps. The high accuracy and reliable positive correlation between simulated and real scan paths illustrate the advantages of the proposed framework in the modelling of attention deployment. We conclude in Chapter 8 with discussions about the potential limitations of the work presented and possible future directions in further pursuing the ViTAL framework. During the course of this project, several parts of this thesis have been published in peer-reviewed academic journals and conference proceedings:

1. Xiao-Peng Hu, Laura Dempere-Marco, Guang-Zhong Yang, "Feature Space Analysis of Visual Search in Medical Image Understanding," Medical Image Perception Conference X, Durham, NC., 11-14 September 2003.
2. Xiao-Peng Hu, Laura Dempere-Marco, Guang-Zhong Yang, "Feature Based Visual Search Analysis in Medical Image Understanding," Proceedings of the 12th European Conference on Eye Movements (ECEM12), Dundee, Scotland, UK, 20-24 August 2003.
3. Laura Dempere-Marco, Xiao-Peng Hu, Guang-Zhong Yang, "Analysis of the Dynamics of Eye Movements in Perceptually Meaningful Feature Spaces," Proceedings of the 12th European Conference on Eye Movements (ECEM12), Dundee. Scotland, UK, 20-24 August 2003.
4. Xiao-Peng Hu, Laura Dempere-Marco, Guang-Zhong Yang, "Hot Spot Detection Based on Feature Space Representation of Visual Search," *IEEE Transactions on Medical Imaging*, vol. 22, no. 9, pp. 1152-1162, 2003.
5. Xiao-Peng Hu, Laura Dempere-Marco, Guang-Zhong Yang, "Hot Spot Detection Based on Feature Space Representation of Visual Search in Medical Imaging," Proceedings of the 4th Annual IEEE-EMBS Information Technology Applications in Biomedicine 2003 (ITAB 2003), pp. 261-264, Birmingham, UK, 24-26 April 2003.
6. Laura Dempere-Marco, Xiao-Peng Hu, Guang-Zhong Yang, "Visual Search in Chest Radiology: Definition of Reference Anatomy for Analysing Visual Search Patterns," Proceedings of the 4th Annual IEEE-EMBS Information Technology Applications in Biomedicine 2003 (ITAB 2003), pp. 187-190, Birmingham, UK, 24-26 April 2003.
7. Laura Dempere-Marco, Xiao-Peng Hu, Guang-Zhong Yang, "Visual Scan-Path Analysis with Feature Space Transient Fixation Moments," Proceedings of SPIE Medical Imaging 2003 Conference, Volume 5034, pp. 503-512, San Diego, California, USA, 15-20 February 2003.
8. Laura Dempere-Marco, Xiao-Peng Hu, Sharyn MacDonald, Stephen Ellis, David Hansell, Guang-Zhong Yang, "The Use of Visual Search for Knowledge Gathering in Image Decision Support," *IEEE Transactions on Medical Imaging*, vol. 21, no. 7, pp. 741-754, 2002.
9. Guang-Zhong Yang, Laura Dempere-Marco, Xiao-Peng Hu, Anthony Rowe, "Visual Search: Psychophysical Models and Practical Applications," *Image and Vision Computing*, vol. 20, pp. 291-305, 2002.
10. Stephen Ellis, Xiao-Peng Hu, Laura Dempere-Marco, Guang-Zhong Yang, David Hansell, "How do radiologists assimilate HRCT images: an assessment based on eye

movements," The Fleischner Society and The European Society of Thoracic Imaging Conference, 2002.

11. Stephen Ellis, Xiao-Peng Hu, Laura Dempere-Marco, Guang-Zhong Yang, David Hansell, "Viewing efficiency of soft-copy reading of high-resolution CT of the lungs," European Radiology, 12:suppl 1, pp.133, 2002.
12. Laura Dempere-Marco, Xiao-Peng Hu, Sharyn MacDonald, Stephen Ellis, David Hansell, Guang-Zhong Yang, "Automatic Feature Learning by Visual Search," Medical Image Perception Conference IX, Warrenton, VA., September 2001.
13. Xiao-Peng Hu, Laura Dempere-Marco, Guang-Zhong Yang, "Implicit Feature Mapping with Visual Search for Diagnostic Decision Support in Medical Imaging," the 11th European Conference on Eye Movements ECEM11, Turku, Finland, August 2001.
14. John Tchalenko, Laura Dempere-Marco, Xiao-Peng Hu, Guang-Zhong Yang, "Quantitative Analysis of Eye Control in Surgical Skill Assessment," the 11th European Conference on Eye Movements ECEM11, Turku, Finland, August 2001.
15. Francois Chabat, Xiao-Peng Hu, David Hansell, Guang-Zhong Yang, "ERS Transform for CT Image Analysis of Small Airways," *IEEE Transactions on Medical Imaging*, vol. 20, no. 9, pp. 942-952, 2001.

# Chapter 2

# Visual Search: A Computational Modelling Perspective

'*The eyes are the windows to the soul.*' This poetical statement originated several centuries ago. It reminds us that the eyes are not just parts of the physical body; they represent an important social cue for humans and other animals -- a cue that is often used to support inferences about the mental states of those whose eyes we observe. By paying attention to where eye gaze has been directed, it is possible to obtain a glimpse of the cognitive process that governs visual search tasks. The nature of how visual search is carried out can be affected in several ways, including our knowledge, interests and expectations. Human eyes do not have a uniform visual response, and the best visual acuity is only within a visual angle of one to two degrees. This is called foveal vision, and for areas that we do not direct our eyes towards when observing a scene, we have to rely on a cruder representation of the objects offered by non-foveal vision, of which the visual acuity drops off dramatically from the centre of focus. In reality, the vast amount of visual information we are exposed to is so complex that we have to absorb and process only a small part of it actively and selectively. When we try to understand a scene, we do not scan it randomly. Instead, we fixate our eyes on particular areas and move between them. This is called 'visual attention', which is used to describe this active and selective activity.

Visual search is a cognitive process involving feature detection, pattern analysis, object identification, reasoning and memory. Extensive research into patterns of eye movements

and brain activity during visual search has been carried out for many years, and a number of computational models have been proposed to illustrate the role of visual attention in visual perception. In this chapter, we will examine the basic concept of visual pathway and techniques for the detection of eye movements and brain activity. Different stages of visual processing and the mechanisms for attention selection will be discussed. We will also outline some of the theoretical frameworks used for the computational modelling of visual search and attention.

# 2.1 Visual Pathway

## 2.1.1 Basic Optical Pathway

For human eyes, the retina forms the inner surface of the eye and is divided into outer pigmented and inner neural layers. The retina can be considered to be anatomically and functionally inside out: the embryological development of the retina is that of out-pouching of the brain that has become sensitive to light. Light must travel through a network of nerves and vessels before reaching the light-sensitive cells. The highest density of light receptors is in the central foveal region, resulting in the greatest resolution of foveal vision. As shown in Figure 2.1, the light receptor cells are linked to ganglion cells *via* bipolar cells, and each bipolar cell receives signals from multiple adjacent light receptor cells and their outputs are then further integrated by ganglion cells from which the final nerve fibres that leave the eye are derived [16]. This hierarchy of nerve cells reduces the input from $120x10^6$ receptors to an output of $1x10^6$ nerve fibres, thereby reducing the required size of the optic nerve 120 fold. There are two types of light-sensitive receptors in the retina: those that respond to different wavelengths of light, and therefore colour, called cones, and those that respond to intensity of light called rods. The macula is an oval shaped condensation of, primarily, cone cells. At the centre of the macula, is a 0.4mm area termed the fovea. It is at the focal point of the lens and contains exclusively cone cells at the greatest concentration in the retina. The density of cone cells gradually declines from the edge of the macula to the retinal periphery where only rod cells are found. The cone density required for detailed colour vision is only found at the fovea. The fixation of an object results in the projection of that object onto the fovea.

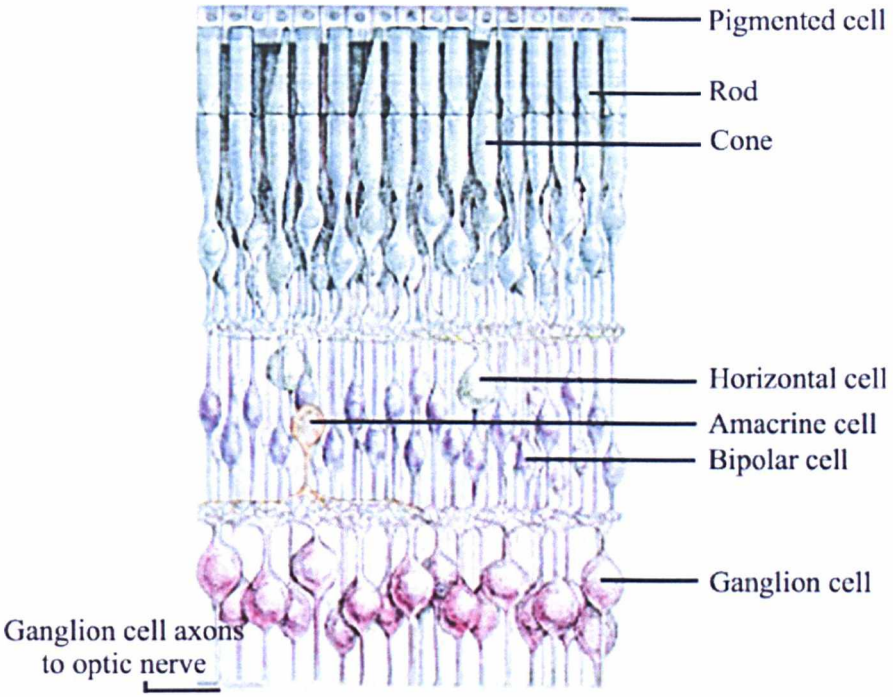



**Figure 2.1** A cross section of the retina (*Adapted from D.H. Hubel's Eye, Brain, and Vision. W.H. Freeman, New York, 1995*).

As shown in Figure 2.2, visual signals generated by photoreceptors are divided at the fovea centre into the temporal and nasal halves, and are taken away by the optic nerve [16]. The fibres from the nasal retinas cross over at the optic chiasm, yet the fibres from the temporal retinas do not cross. This means that the visual information from the left visual field is transmitted to the right hemisphere, and the information from the right field to the left hemisphere. The visual signal is carried further by the optic tract to the lateral geniculate nucleus (LGN) of the thalamus. Nuclei in the LGN are arranged on six layers. Layers 1 and 2 receive information about movement and contrast. Layers 3 to 6 receive information about shape and colour. From the LGN, the fibres spread out through the optic radiations and ultimately travel to the primary visual cortex (striate cortex or V1) at the back of the brain.

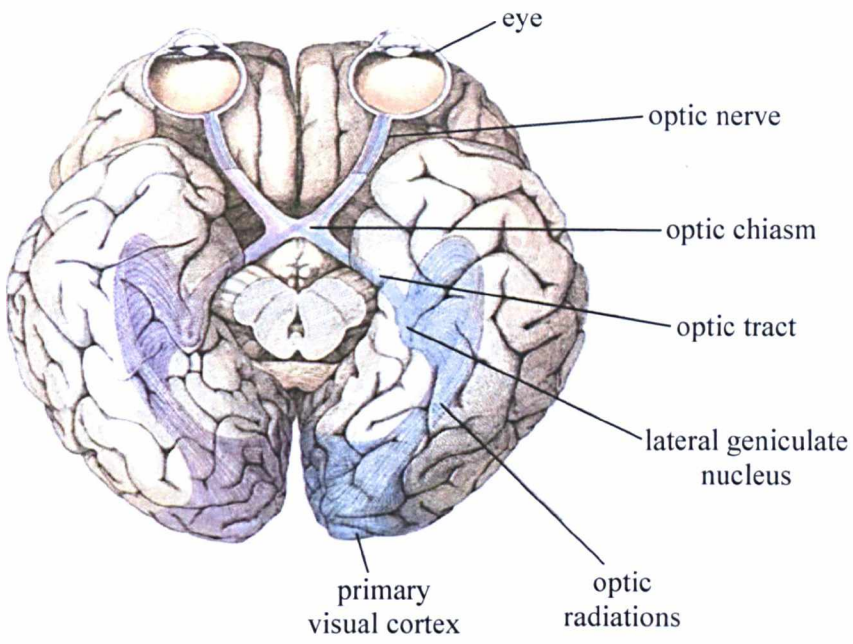



**Figure 2.2** A schematic illustration of the basic optical pathway (*Adapted from D.H. Hubel's Eye, Brain, and Vision. W.H. Freeman, New York, 1995*).

The first human cerebral cortical area for visual information processing is V1, which performs primary pattern-analysing tasks [17]. It is estimated that 25-40% of the human cerebral cortex is associated with visual information processing, and about 10-20% of that volume is V1. The neural cells in the primary visual cortex can be divided into two major groups: simple cells and complex cells [16, 18, 19]. Both of these cells are selective to luminance lines of a particular orientation. Simple cells respond to width and position of stimulus, whereas complex cells, which occupy most of the population of cells in V1, are particularly responsive to moving direction. The cells responsive to particular characteristics of visual stimulus are also called feature detectors.

## 2.1.2 Ventral and Dorsal Visual Streams

After the primary visual cortex, visual information diverges into two streams: the dorsal stream and the ventral stream [20-22], as schematically illustrated in Figure 2.3. The dorsal stream (or 'where' stream), from the primary visual cortex (and the superior colliculus *via* the pulvinar) to the posterior parietal cortex, is involved in action control including visual spatial localisation and visuomotor control, whereas the ventral stream (or 'what' stream), from the primary visual cortex to the infero-temporal cortex, is responsible for visual perception including object recognition and identification [21, 23].

This theory has been bolstered by numerous sources of psychophysical evidence. For instance, lesion studies of human and animal brains demonstrate that dorsal lesion would cause optical ataxia -- patients with optical ataxia can recognise objects but cannot grasp or reach the objects by using vision. Ventral lesion damage would result in visual agnosia -- patients with visual agnosia can describe the visual features of objects but cannot identify them by using vision.

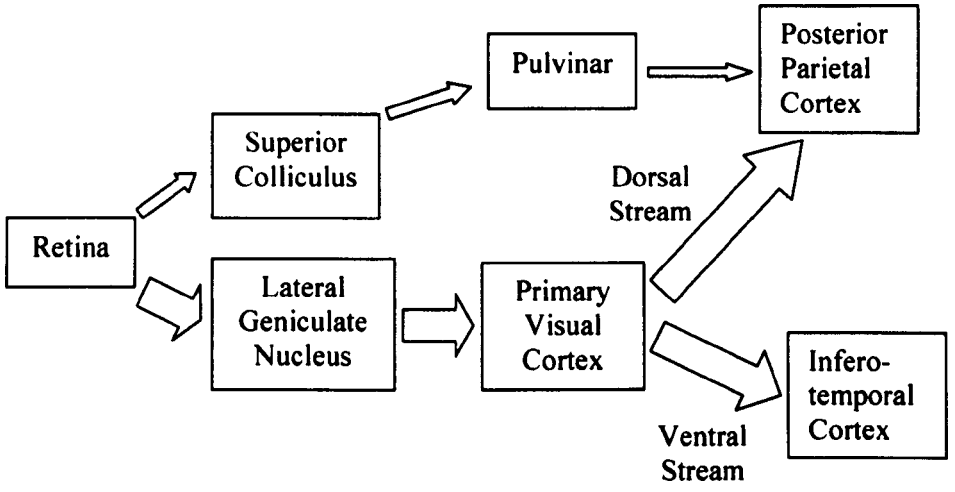



**Figure 2.3** A schematic illustration of the dorsal and ventral visual streams.

There are intrinsic relationships between specific areas in the ventral and dorsal streams and their functions. The following is a list of common findings that have been documented in neuroscience experiments [20-22]:

***Ventral stream***

*V4 (Fourth Visual Area): colour/simple shapes*

*LO/LOC (Lateral Occipital Complex): object and shape processing.*

*FFA (Fusiform Face Area): face processing.*

*PPA (Parahippocampal Place Area): place processing (scenes, houses).*

*EBA (Extrastriate Body Area): human body processing.*

***Dorsal stream***

*V5/MT (Middle Temporal Area): movement analysis.*

*VIP (Ventral Intraparietal Area): multimodal motion responses (visual, auditory, tactile).*

*LIP (Lateral Intraparietal Area): saccadic eye movement processing.*

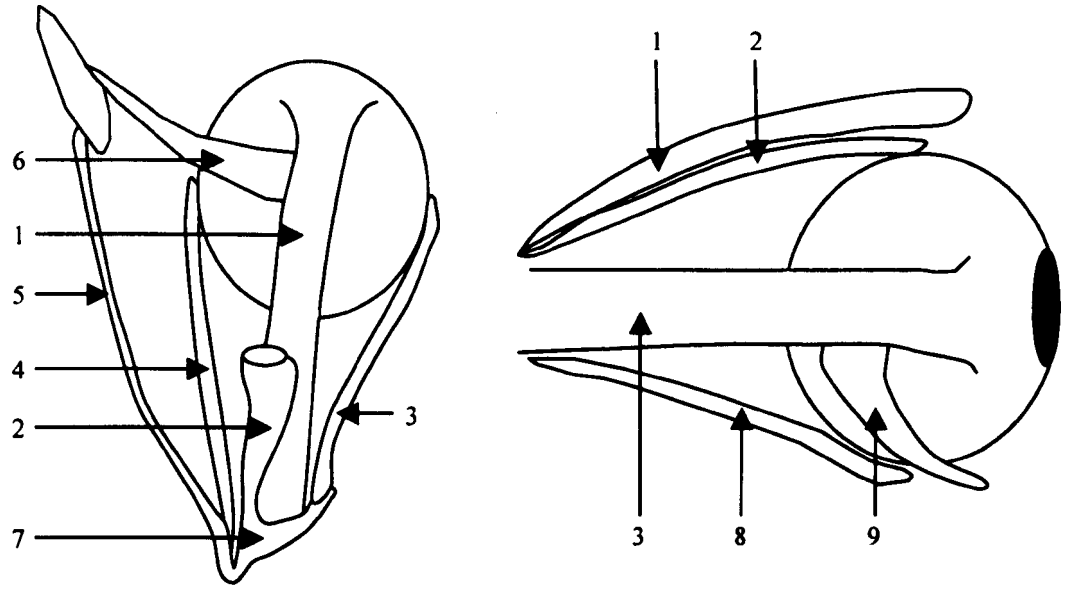
*SPL (Superior Parietal Lobule): attentional responses.*

The interaction between the ventral and dorsal visual streams is important for both 'what' and 'where' activities. However, the causal relationship between the two streams is still not well understood. It has been suggested that the cognitive communication between them is controlled by a distributed network, which consists of several parts of the brain cortex [24, 25]. Recent research indicates that the prefrontal cortex plays a key role in this cognitive control [26-28].

### 2.1.3 Eye Movement Control

Anatomically, each eye has six extrinsic eye muscles that originate at the bony orbit and insert into the outer surface of the globe, allowing the eyes to follow a moving object and provide external support that help to maintain the shape of the eyeball and hold it in the orbit. Four of the muscles originate at a common tendinous ring, the annular ring at the back of the orbit, and run directly to their insertion on the globe to control the superior/inferior and lateral movements of the eye. The remaining two muscles are the superior and inferior oblique muscles. The superior oblique muscle rotates the eye downward and laterally; it originates in common with the rectus muscles and runs along the medial wall of the orbit and inserts on the superolateral aspect of the globe having passed through a fibrocartilaginous loop. The inferior oblique muscle rotates the eye up and laterally; it originates from the medial orbit surface and runs lateral and obliquely to insert on the inferolateral eye surface. Figure 2.4 illustrates the extrinsic muscles of the eye [29].

The abducens and trochlear nerves innervate the lateral rectus muscle and the superior oblique muscle respectively. The oculomotor nerves innervate the remaining extrinsic eye muscles. The need for six muscles when four should be sufficient is a result of the mechanics of the inferior and superior recti, which rotate the eye medially while elevating or depressing it. The extra, oblique muscles counteract this effect. Eye movements are usually synchronous, each eye fixating the same object. However, if an object nearer to the eye is fixated, the eyes converge and conversely diverge when fixating objects are more distant.




(a) view from above (b) lateral view

**Figure 2.4** A diagrammatic overview of the extrinsic muscles of the eye. The *medial* and *lateral recti* control sideway movements; the *superior* and *inferior recti* control up/down movements; the *superior* and *inferior obliques* control twist movements. 1: superior rectus; 2: levator palbebrae superioris; 3: lateral rectus; 4: medial rectus; 5: superior oblique; 6: reflected tendon of the superior oblique; 7: annulus of Zinn; 8: inferior rectus; 9: inferior oblique (*Adapted from H. Davson's Physiology of the Eye. 4th ed, Academic Press, New York, 1980*).

During visual recognition, the superior colliculus (SC) of the middle brain receives inputs from several cortical regions (the temporal cortex and parietal cortex) to form signals for eye movement control. The rostral pole region of the superior colliculus is crucial in fixation and saccade control [30, 31]. The SC also receives direct inputs from the retina *via* a visual pathway running through the photoreceptors (rods and cones), optic nerve, optic chiasm and optic tract to the SC [22]. These direct inputs are involved in the formation of some visual reflexes such as the protective action of eye closure to stimuli.

The intrinsic dynamics of eye movement are complex, but the following movements are common. They include saccadic, miniature, pursuit, compensatory, vergence and optokinetic movements. Voluntary fixation allows the eyes to move to find the object that is to be fixated on; involuntary fixation holds the eye such that the fixated object remains within foveal vision. Fixation points are the delimiters of saccadic movements. It is thought that a fixation time of at least 100 milliseconds is required to allow high-resolution perception from the macula. Voluntary fixation movements are controlled by a small cortical field in the pre-motor cortical regions of the frontal lobe, distant from the visual cortex and brainstem, and override the involuntary fixation movements when an alternate fixation point is to be found. Voluntary fixation movements are saccadic; jerky

movements that quickly move the eye from one fixation point to another, allowing rapid foveal interrogation of multiple sites in the field of view. Saccadic movements occupy only 10 per cent of the total time spent in moving the eyes; the remaining 90 percent is allocated to fixation movements. During saccadic movement the visual signals from the retina are suppressed so that the movement is not perceived visually, resulting in a stable field of view [32].

Involuntary fixation is regulated by a negative feedback mechanism that prevents the object of attention from leaving the foveal portion of the retina. The eyes normally have three types of continuous but almost imperceptible movements: (1) continuous tremor at a rate of 30 to 80 Hz, caused by successive contractions of motor units in the ocular muscles; (2) a slow drift of the eyeballs; and (3) sudden flicking movements controlled by the involuntary fixation mechanism. Tremor movements do not usually affect fixation, but the drifting movement can move the object of fixation outside the foveal region, at which point there is an automatic flicking movement that returns the fixated object's image to the centre of the fovea [33].

Another common movement of the eyes is pursuit movement, which enables the eye to remain fixed on a moving object. A highly developed cortical mechanism automatically detects the course of movement of an object and then gradually develops a similar course of movement for the eyes. The longer the time of pursuit, the smoother the pursuit movement becomes.

### 2.1.4 Visual Acuity

Visual acuity refers to the spatial resolving capacity of the visual system. It is mainly limited by the PSF (Point Spread Function) due to distortions created by the optics of the eye and the spatial density of the mosaic of the photoreceptors [34]. Other factors affecting visual acuity include refractive error, illumination, the exposure time of the target, the area of the retina stimulated, and eye movements. There are various ways to measure and specify visual acuity depending on the type of acuity tasks, which include target detection, recognition and localisation [35]. Target detection requires only the perception of the presence or absence of an aspect of the stimuli; target recognition requires the identification or naming of a target, such as the Snellen letters [36]. Target localisation involves discriminating spatial differences of segments of a stimulus, such as a break or discontinuity in contour. The Snellen letters are designed so that the overall

letter height is equal to five times the size of the critical detail (stroke width and gap width). The Snellen acuity is determined according to the smallest letters in the chart that the observer can correctly identify.

Studies on the photoreceptor density show that the cone spacing at the fovea is approximately 28 seconds of arc, which results in about 60 cycles per degree of maximum visual resolution. As indicated in Figure 2.5 [37], the foveola, which is a rod-free area within the fovea, contains 147,000 cones/mm$^2$. At about 10 degrees, the number of the cones drops sharply to less than 20,000 cones/mm$^2$. Due to the densely packed cones at the fovea, visual acuity reaches its peak at the centre of fixation. Figure 2.6 indicates the visual acuity at various eccentricities [35]. At a distance of 5 minutes of arc from the centre of fixation, there is a measurable loss in visual acuity. At 10 minutes of arc (one-sixth of a degree) from fixation, there is a 25 % loss of visual acuity.

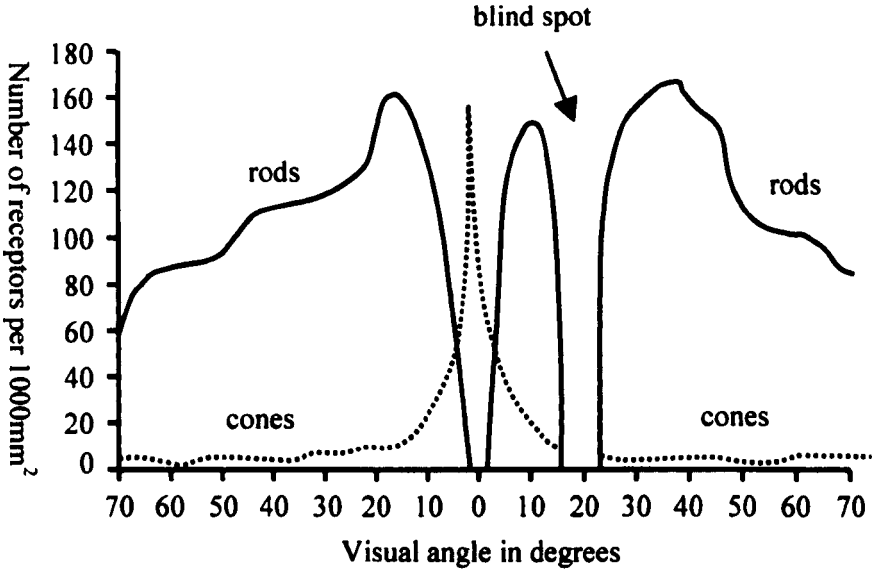



**Figure 2.5** A distribution of cones (dotted lines) and rods (solid lines) on the human retina (*Adapted from R.N. Haber and M. Hershenson's The Psychology of Visual Perception. Holt, Rinehart, and Winston, Inc., New York, 1973*).

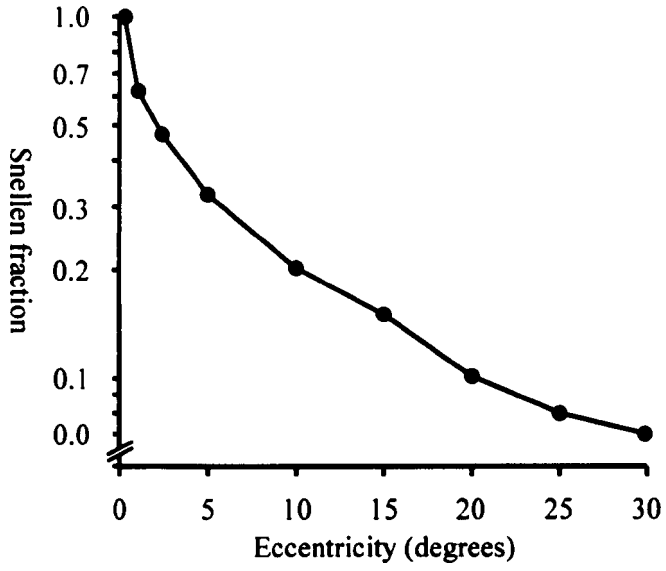



**Figure 2.6** The effect of eccentricity on visual acuity (*Adapted from G. Westheimer's Visual Acuity. Chapter 17. In: Moses, R. A. and Hart, W. M. (ed) Adler's Physiology of the eye, Clinical Application. St. Louis: The C.V. Mosby Company, 1987*).

The above effect can be explained by using the sampling theory of visual resolution [38, 39]. This theory indicates that visual acuity is closely related to the density of the photoreceptors. The quantitative relationship between visual acuity $V$ and cell density $D$ is:

$$V = 0.5 * \sqrt{D} \tag{2.1}$$

When the visual acuity predicted from the cone packing density is compared to the visual acuity from the visual acuity experiment, close correlation is evident up to approximately two degrees eccentricity [40]. However, at larger eccentricities, visual acuity is worse than that predicted by cone spacing, which may indicate that the mechanism involved is not restricted to the retinal level [41]. It has been reported [42, 43] that eye micro-movements have little effect on the performance in the central retina since visual acuity is mainly limited by optical blurring. However, eye micro-movements clearly improve the performance in the retinal periphery since they reduce the effect of aliasing due to neural under-sampling.

## 2.2 Experiments for Measuring Visual Search Responses

To elucidate the intrinsic links between visual search and the underlying cognitive processes, a number of experiments have been conducted in psychology and neuroscience. Two common examples are the measurement of Response Time (RT) and visual search accuracies. They are usually studied as a function of the search set size. The RT and accuracy test provides a basic paradigm for determining the serial/parallel dichotomy in visual search [1]. For example, it is easier to find 'X' than 'L' in Figure 2.7. Accordingly, the plot of the corresponding RT/set-size function of searching for 'X' will be relatively flat. It has been demonstrated that basic (primitive) features such as colour, orientation, curvature, size and motion can influence the 'pop-out' attribute of an object.

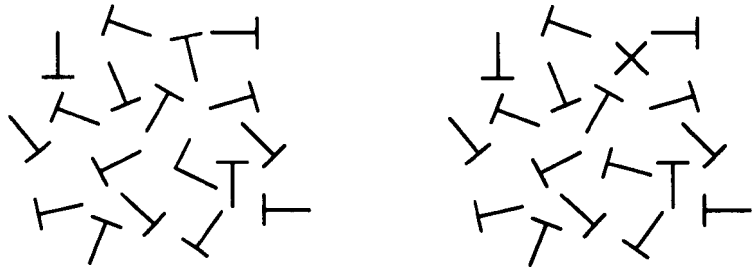


**Figure 2.7** Patterns to illustrate the serial/parallel dichotomy in visual search. It is easier to find 'X' in the right pattern than 'L' in the left pattern.

In practice, visual search tasks require the identification of multiple visual cues leading to the target. Cueing and priming have been influential paradigms in the study of visual search. These methods change visual search experiments from a passive mode to an active one. The efficiency of a visual search process can be altered by using cueing and priming methods to change observers' expectation (valid, invalid or neutral) or provide observers with some prior knowledge and experience [44, 45]. In a 'cueing task', the location of a visual stimulus is indicated with a pre-cue. Typically, a valid cue leads to shorter response time and higher accuracy. On the other hand, an invalid cue results in a slow and inaccurate response. Priming changes an activation rate or expectation for a visual feature or object. Similar to the cueing effect, a primed feature, pattern or object is usually recognised faster and more accurately than unprimed stimuli. Cueing and priming paradigms have been utilised to detect top-down control activity of the brain by analysing the pattern of brain activity during visual search processes [46]. They have also been applied to test the relationship between eye movements and visual attention [44, 45] as well as the inhibition of return [47].

Imagery experiments are also common in the studies of psychology and neuroscience. In a mental imagery experiment conducted by O'Craven *et al* [48-50], patterns of brain activity are recorded when subjects are asked to imagine the shape of an object, *e.g.* a face and a house, according to their knowledge and experience without showing the object. This pattern is then compared to the one recorded when subjects are shown the object in order to analyse the top-down control and visual memory [51].

The techniques used for the detection of eye movements and brain activity during visual search have evolved over years. Objective and direct measures provided by physical techniques are crucial for avoiding subjective judgment made in certain behaviour studies. For instance, we cannot make a detailed judgment, such as whether a visual stimulus is ignored by pre-attentive detection or abandoned by attention selection, without physiological measures of neural responses by using functional brain-imaging techniques.

### 2.2.1 Positron Emission Tomography (PET)

PET scanning is a technique for detecting changes of the cerebral blood flow associated with brain functions in regions of the brain. A small amount of radioactive glucose tracer is injected into the bloodstream in advance. Positively charged positrons are released during the decay of the nuclei in the injected tracer. When a positron collides with a negatively charged electron, both are annihilated and two photons (or gamma rays) are emitted. The emitted photons move in opposite directions and are detected by the detector ring of the PET scanner. Active areas of the brain consume more glucose, therefore resulting in a stronger signal, which is then converted by a computer into a picture of brain activity. This technique, however, is restricted by its low spatial and temporal resolution and the invasive injection of the radiation material.

### 2.2.2 Event-Related Potentials (ERP)

The principal application of ERP analysis is to determine the time course of higher-level processes in the human brain. ERP captures brain responses to stimulus events by recording electrical voltage waves at human scalp sites. Since the event-related electrical signal is often affected by electrical signals generated by unrelated brain activity, it is

necessary to average the recordings from repeated presentations of the stimulus. The higher temporal resolution of ERP (compared to fMRI) enables it to provide the timing information of neural processes. However, the limited spatial resolution restricts its ability to visualise activity in the neural structures.

### 2.2.3 Functional Magnetic Resonance Imaging (fMRI)

Functional MRI allows the mapping of the functions of the various regions of the human brain by using MRI technology to detect quick and tiny metabolic changes in particular areas of the brain. For example, BOLD-fMRI determines the brain activity by detecting Blood-Oxygen-Level-Dependent (BOLD) contrast [52-54]. Oxygen is carried by haemoglobin in blood. Haemoglobin with an oxygen molecule bound to it is called oxyhaemoglobin and haemoglobin without oxygen is called deoxyhaemoglobin. Deoxyhaemoglobin is a paramagnetic molecule whereas oxyhaemoglobin is diamagnetic. Neural activity requires oxygen, which results in a larger increase in cerebral blood flow to bring more oxyhaemoglobin and, consequently, a local decrease in deoxyhaemoglobin. This effect gives rise to an increase of the MR signal in activated regions. fMRI has a higher temporal and spatial resolution than PET [55]. Associated with event-related techniques, it can be used to visualise brain activity in response to specific stimuli.

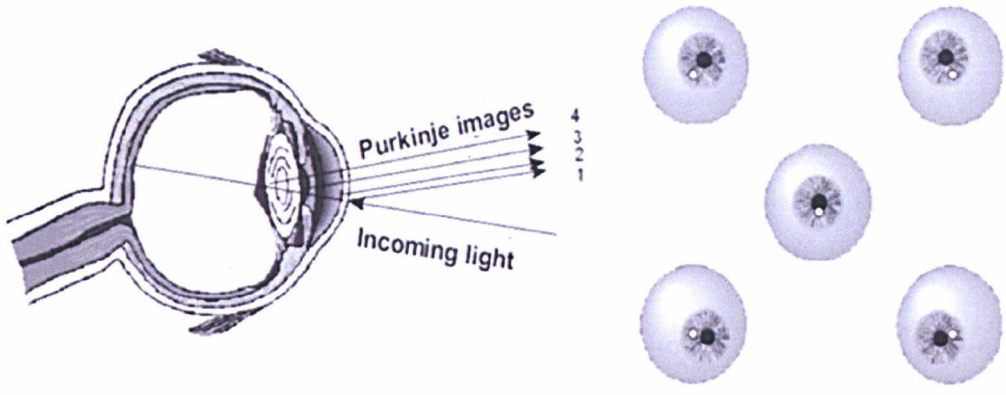
### 2.2.4 Disruption Techniques (Lesion, Damage, Disorder)

The techniques described above are able to determine brain activity in response to stimuli, but cannot determine the necessity of brain activity in processing visual stimulus. Disruption techniques, such as analysis of brain damage and disorder, have been used to establish the detailed functional picture of the brain cortex. In fact, analysing the effects of damage to certain parts of the brain is a traditional technique for studying brain functions. For example, the subtractive model attempts to determine the relationship between brain structures and behavioural capacities by analysing the behavioural changes that follow brain damage. Although the tendency of this model to assign functions to specific brain structures has frequently been criticised, it is still widely used because it is useful for establishing correlations between brain structures and behaviours.

## 2.3 Eye Tracking

In general, there are three types of eye trackers used to measure eye movements [56]: Electro-Oculography (EOG), Scleral Contact Lens/Search Coil, and Video-(Photo-) Oculography (POG, VOG). EOG records the skin's electric potential differences to measure eye movements by attaching electrodes around the eye. Scleral Contact Lens/Search Coil relies on mounting a contact lens directly on the eye. The intrusive nature of EOG and Scleral Contact Lens/Search Coil techniques constrains their usability. On the other hand, POG and VOG techniques measure eye movements by detecting visible features of the eye, such as the boundary of the pupil and corneal reflection. POG and VOG techniques are now widely applied because they are non-intrusive, accurate and allow minor head movement.

Although a video-based eye-tracking apparatus can be table-mounted or head-mounted, the optic principles of both systems are identical with the exception of size. The relative position between the centre of the pupil and the cornea changes with eye rotation, but remains relatively constant with minor head movements. Therefore, the viewer's point of regard can be measured despite minor head movements after appropriate calibration procedures. Given a light source, typically infra-red, four Purkinje reflections or Purkinje images are formed due to the structure of the eyes [57]. The first and second reflections are from the front and rear surface of the cornea, and the third and fourth reflections are from the front and rear surfaces of the lens. Figure 2.8 demonstrates how the four Purkinje reflections are formed as well as the relative positions of the pupil and the first Purkinje reflection, when the eye fixates to the centre and the four corners of a computer screen [4]. The *Dual-Purkinje Image* technique uses the first and fourth Purkinje images for tracking the direction of gaze. It is generally more accurate than other eye-tracking techniques. However, since the fourth Purkinje image is rather weak, the surrounding lighting must be strictly controlled.




(a) (b)

**Figure 2.8** (a) The basic principle of how the four Purkinje images are formed. (b) The relative positions of the pupil and the first Purkinje reflection during visual search as seen from the camera when the observer fixates to the centre and the four corners of the screen. (*Adapted from Visual search: psychophysical models and practical applications, by G.Z. Yang, L. Dempere-Marco, X.P. Hu and A. Rowe, Image and Vision Computing, vol. 20, pp. 291-305, 2002*)

## 2.4 Stages of Visual Processing

Visual processing involves cognitive activities including sensation, perception and memory. It is now generally accepted that visual processing can be separated into pre-attentive and attentive processing stages. Several parts of the brain are involved in the processing of visual information and therefore the time course of visual processing is helpful for understanding the underlying cognitive mechanisms.

### 2.4.1 Pre-attentive and Attentive Processing

Pre-attentive vision occurs before visual attention. It is a fast and parallel operation over the entire visual field with unlimited capacity. Following pre-attentive vision, it comes the attentive vision. In contrast to pre-attentive vision, it operates slowly and serially over a small portion of the visual fields with limited capacity. The term to describe the dichotomy of these two stages is 'visual attention', which stems from the limited ability of our biological visual system. To avoid capacity overflow, the limited cognitive resources need to be allocated selectively towards the processing of important information in a serial manner.

According to Feature Integration Theory (FIT) [2], pre-attentive features over the entire visual fields are independently coded on the feature map in parallel at the pre-attentive stage. During the attentive stage, visual attention is necessary to search serially through the master map of location when information about the relationship between features is required. FIT can nicely explain the 'pop-out' attribute of pre-attentive features and the linearity of the RT/set-size function of feature conjunction search. However, this model is unable to account for the continuous gradients of the RT/set-size function of parallel conjunction search [1].

## 2.4.2 The time Course for Visual Processing

The human visual system appears to be able to perform highly demanding object-detection tasks in natural scenes very quickly. The ERP analysis of a go/no-go categorisation task experiment [58] illustrates that the latency of visual signals from the primary visual cortex to the infero-temporal cortex is within 150 ms, at least 50 ms [59] of which are spent on visual stimuli sensation, such as edge detection and motion detection.

Table 2.1 shows a time course for the rapid visual processing [58-60] of visual information. More than 10 different cortical regions are involved in visual information processing, which can be completed in about 150 ms [58]. Such a high processing speed presents a challenge for neuroscientists to interpret how, intrinsically, everything happens.

**Table 2.1** A representative time course for rapid visual processing in the brain.

| Process | Areas | Latency (monkey) |
|---|---|---|
| Early sensory processes | V1/V2 | 40-70ms |
| Intermediate processes (grouping) | V4 | 60-80ms |
| Object-level processing | IT | 80-100ms |
| High-order judgments (decisions) | PFC | 100-130ms |

## 2.4.3 Attention Selection

Visual attention is a selective activity of visual search, which allows the decomposition of computationally demanding visual analysis into less difficult small tasks. Physiological research demonstrates that visual attention affects neural responses to visual stimuli by

enhancing the activity in stimulated visual cortical regions and suppressing the activity in unstimulated visual cortical regions *via* (additive) baseline shifts, (multiplicative) gain modulation or both [61-66]. Attention is also believed to play an important role in solving the 'binding problem' [2, 67], *i.e.* how to bind information, coded in a distributed manner, about various visual features of an object into an integrated perceptual representation of the object. Attention selection is crucial for avoiding the overloading of cognitive processing. The locus of the selection is the intersection before which the processing resources are unlimited and after which the resources are limited.

### *1) Early (Passive) Selection*

Early work on attention by Broadbent and Treisman resulted in the Early Filtering Model [68] and the Attenuation Model [69]. These early selection models suggested that whole stimuli are analysed by pre-attentive detection, and object recognition is done at a high level. Attention selection occurs after pre-attentive detection and before object recognition. According to the Early Filtering Model, the processing capability of the brain is limited; therefore only relevant information is selected for further processing and irrelevant materials need to be filtered out by an information bottleneck mechanism based upon the physical properties of stimuli. On the other hand, Treisman's Attenuation Theory suggests that unattended information might pass through the selective threshold rather than being completely eliminated. The volume (or threshold) of an information channel can be turned down or attenuated in response to the salience or expectation of the information to be processed.

### *2) Later (Active) Selection*

The Later Selection Model [70] proposed by Deutsch suggested that attention selection occurs after stimuli have been fully identified. This model assumes that all the information, both relevant and irrelevant, can enter the short-term memory in order to be recognised, but material dies rapidly if it is not attended. The role of attention is to select meaningful information that needs to be remembered for further processing, rather than protecting the limited capacity of the perceptual system, as argued by Broadbent.

Both the Early and Later Selection models agree that early visual processing is not affected by attention selection. This is supported by ERP experiments [71] in which responses to attended and unattended stimuli are similar in the primary visual cortex, but

differ in the later stages of cortical visual processing in both the ventral and the dorsal streams. On the other hand, several recent fMRI studies [61, 72-74] indicate attentional modulations in the primary visual cortex and these are thought to be related to early stage visual processing. Furthermore, cortical activity is also found to be associated with object recognition [75, 76] and processing load [77-79]. Early selection seems to fit in with high workload situations and late selection with low workload circumstances. From a psychophysical point of view, visual selection can happen at every stage of visual processing, although the separation of later or early selection is useful in the analysis of visual search behaviours.

### 2.4.4 Space, Feature and Object-Based Selection

Psychological and physiological experiments indicate that attention selection can be performed in the spatial domain [44, 45, 80], *i.e.* the selection is allocated to spatial locations; in feature space, *i.e.* the selection is allocated to visual features [2, 67]; or in object space, *i.e.* the selection is allocated to the presentations of all the features that define an object [75, 76, 81]. Ponser [44, 45] used the 'mental spotlight' metaphor to depict internal mental attention searching in the spatial domain. Cueing task experiments showed that cued locations are easier to access shortly after the cue but are more difficult to access later (inhibition of return) [47]. There is also supportive evidence for space-based attention from physiological research [61, 62, 80]. However, failure to replicate the inhibition of return was reported in an experiment conducted by Wolfe [82]. Evidence for feature-based and object-based selection was found in psychological and physiological experiments [63, 83] in which subjects were presented with spatially inseparable objects with different features. In an fMRI experiment conducted by O'Craven *et al* [75], visual stimuli (face, house and motion) were transparently superimposed and displayed in the same location. Enhancement of neural responses to these stimuli was detected in the cortical regions such as the fusiform face area (FFA) and the middle temporal area (MT), indicating feature-based or object-based (rather than location-based) attention selection.

Space-based, feature-based and object-based attention may not be mutually exclusive. These selections might work serially or simultaneously in different cognitive stages [84, 85] and according to the processing load [77-79]. At the early stage or when the processing load is high, the visual system seems to work at the level of locations and

visual features. Object-based selection usually occurs at late stages and when the working load is low.

### 2.4.5 Attention Deployment and Eye Movement

Eye movements consist of a sequence of saccades and fixations. Saccades are fast ballistic eye movements from one fixation to another, whereas the eye has small draft movements during fixations [30]. Information is largely gathered during fixations while very little information is gained during saccades. Fixation duration and saccadic landing position are two important measures for scan path analysis. The typical duration of a fixation is about 200-250ms. Figure 2.9 shows a scan-path of the eye when an observer was asked to search for a pair of father and son in the picture.



**Figure 2.9** A typical scan-path recorded by an eye-tracker. The search task is to identify a pair of father and son. The size of circles represents the dwell time of the fixations whereas the lines between fixations indicate the saccadic eye movements.

Saccades are usually directed to regions sharing common features with the target during visual search. Consequently, the position and duration of fixations are important in the analysis of eye-scanning behaviours. The following are some results from human perception and performance investigations.

- *Information sampling*: Information is selected actively according to interest. The cost of sampling is usually proportional to the complexity of an image to be interpreted.

There could be over-sampling of a part of an image if infrequent information appears [86, 87].

- *Sampling cost*: the cost of sampling is reduced if the points to be sampled are spatially close. This means that subjects may minimise eye movements; they tend to make most comparisons between spatially adjacent data [88].
- *Scan pattern*: some subjects exhibit systematic scan patterns, such as left-to-right and top-to-bottom scanning, even if these have no functional relevance to the task [89].

The relationship between overt eye movement and covert visual attention has been investigated for many years. Existing work by Posner in his cueing paradigm [44, 45] has established that attention can be shifted covertly to different locations without any corresponding eye movements. It is estimated that the speed of covert attention movements could be 30 shifts per second [2, 30], much faster than that of eye movements. Accordingly, there could be about 8-10 attention shifts during an eye fixation (200-250ms). It is also reported that overt eye movement and covert visual attention are very closely related. Although attention can be moved without eye movement, the eye cannot be moved without attention shifts. A reading study by Morrison suggests that covert attention could be deployed first to a position in preparation for the following saccade [90]. The mechanism of covert attention and its relationship with eye movements is still not clearly understood.

## 2.5 Information Representation

Visual attention is directed by two types of information: bottom-up or image-based information and top-down or task-dependent information [1]. Our visual system first breaks visual stimuli into small components or patterns and, after processing it combines these components back together into an integrated and more abstract representation. It is suggested that these breaking and combining procedures are very similar to the forward and backward Fourier transform and, to some degree, can be regarded as linear [91, 92].

### 2.5.1 Bottom-Up (Image-Based, Task-Independent) Information

Bottom-up information refers to visual sensory stimulation, which is independent of visual search tasks. As indicated by psychological experiments [2, 67], some features such as colour and orientation are so salient to our visual system that they 'pop-out' of

visual scenes almost without any apparent visual computational cost, but some others are not so easy to detect.

The specialised receptive fields of neurons in the primary visual cortex (area V1) were found by Hubel and Wiesel in their study on the vision of frogs and cats [16-18]. Campbell and Robson subsequently reported that visual neurons are selectively sensitive to local light patterns [92]. Brain-imaging research also indicates that specific cortical regions respond to specific views of complex objects. For instance, neural activity in the Fusiform Face Area (FFA) is higher when subjects attend to faces [93], and in the Parahippocampal Place Area (PPA) when subjects attend to houses [94]. These discoveries provide neural evidence for theories of distributed bottom-up representation in the brain [95].

### 2.5.2 Top-Down (Task-Dependent) Information

Top-down information represents prior knowledge rather than sensory stimulation. Top-down activity plays an important role in visual search and is regarded as a voluntary 'effort' for the visual system to participate actively in visual search processes [96].

Similar to bottom-up information, top-down information is also coded in a distributed fashion, *i.e.* the specialised cortical regions are involved in specific visual information. As demonstrated in several experiments, increased neural activity (baseline shifts) is detected in specific cortical regions, such as the frontal, parietal and primary visual cortices, when subjects expect a coming visual target [46] or imagine a specific visual object [48-50]. In order to distinguish top-down neural activity from bottom-up neural activity in these experiments, neural responses to visual stimuli were compared with neural responses to instructive cues [97], and neural activities recorded when subjects directed attention covertly to an expected target location were compared with neural activities recorded when subjects maintained fixation [46].

The results of these studies indicate that the top-down control signal originates mainly from a network in the frontal and parietal cortices and is sent back to the early visual cortex. Damage in the regions of the parietal and frontal lobes often leads to spatial neglect [98, 99]. For instance, patients with unilateral neglect often attend to only one side of objects in the scene [100]. This symptom also suggests that the frontal and parietal cortices are involved in top-down control.

### 2.5.3 Visual Memory

Recent findings from brain imaging, lesion studies and visual mental imagery assessment provide evidence for the interpretation of visual memory. It has been reported [51] that specific regions of the frontal cortex seem to be associated with strategic aspects of visual memory, *e.g.* retrieval attempt and monitoring. The parietal and frontal regions provide signals to indicate that retrieved information is from the memory. Cortex regions associated with sensory processing are also associated with the corresponding memory processes. Thus, memory contents are very likely to be coded in a distributed network.

Behavioural research on visual memory demonstrates that changes of scene details are often overlooked by observers, especially when these changes occur during a saccade. This so-called *change blindness* phenomenon stresses the importance of visual attention in visual memory, and suggests that the internal coding of visual information is sparse and only focuses on certain salient features [101-105].

Psychological experiments also show that the capacity of short-term visual memory is limited. In an experiment carried out by Horowize and Wolfe [106], no significant difference in visual search efficiency was recorded when a search pattern was randomly rearranged a short time after subjects searched for a target on it, compared with when the search pattern remained unchanged. Accordingly, they concluded that 'visual search has no memory'. The limitation of short-term visual memory can lead to the over-sampling of parts of the image if infrequent information appears [86]. Strategic scanning is often used in visual search to compensate for this limitation [107].

### 2.5.4 The Binding Problem

The binding problem arises from the distributed representation of visual information: how information (coded in several separate regions of the brain) is integrated to represent an object. The core of FIT [2] is to use spatial attention to solve the binding problem. Features at the attended location are bound together. Accordingly, only one object can be coded at a time on the basis of its location. Recent studies have extensively investigated the function of the prefrontal cortex [26-28]. The results indicate that this region is involved in constructing and representing the relationship of a diverse range of

information, and therefore is central to the process of synthesising the functions of the dorsal and ventral streams [27, 28, 108].

### 2.5.5 Object Recognition

Object recognition is one of the most difficult tasks for vision. Identification and categorisation are the two main tasks of recognition and are closely related [109]. Two types of object recognition theories have been proposed: image-based (or view-based) theory and structure-based (or object-centred) theory. Image-based theory uses the image (or view-specific features) as the representation of objects. Accordingly, object identification involves matching the visual image with the object image in memory [110, 111]. On the other hand, structure-based theory separates an object to be recognised into primitive local components, each of which has some special geometrical features. Object recognition is realised by analysing the relationship between these primitive components [112]. Image-based theory is supported by evidence from psychophysical and physiological research. Brain imaging and lesion studies indicate that several distinct and independent neural pathways are involved in object recognition, and cells in specific cortical regions respond to specific views of an object. For instance, neurons in the fusiform face area (FFA) are sensitive to faces [93] and neurons in the parahippocampal place area (PPA) are sensitive to houses [94]. Object recognition involves both bottom-up activities, such as sensory processes, and top-down activities, such as memory activation and attention. Studies of mental imagery and expectation indicate that neural activities recorded when subjects were imagining or expecting a stimulus were similar to activities recorded when the stimulus was presented [46, 48-50]. These similarities strongly suggest the involvement of top-down activity in object recognition.

Figure 2.10 illustrates a feed-forward model for object recognition [109]. Feature units extract visual features, which then serve as the input to view-tuned units. These units are tuned to full or partial views of objects. On top of the view-tuned units is the object-tuned unit layer. The units in this layer are tuned to tolerate object-dependent transformations such as rotation in depth. Object categorisation and identification are based on the outputs of both view-tuned units and object-tuned units. The feature units in this framework may be located in V1 and the posterior infero-temporal cortex (PIT). The view- and object-tuned units may be located in the anterior infero-temporal cortex (AIT); and object categorisation and identification may happen in the prefrontal cortex (PFC) and some parts of the infero-temporal cortex (IT).

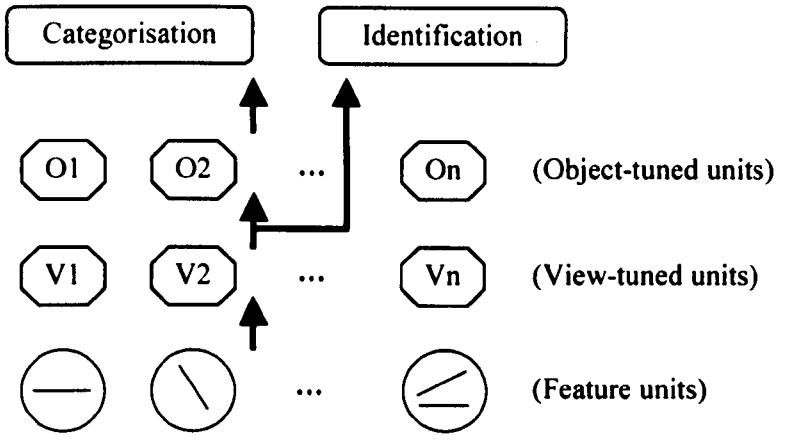



**Figure 2.10** A feed-forward model for object recognition.

## 2.6 Computational Modelling and Simulation

Computational modelling is a tool for the description and simulation of cognitive processes in a formal way. It allows one to think concretely about these processes and to be explicit about the underlying mechanisms. Computational models can be divided into three categories: biologically plausible models, mathematically plausible models and mixed models. The biologically plausible approaches try to explain the cognitive mechanisms by simulating activities in the human brain during visual search, whereas the mathematically plausible approaches attempt to build mathematically accurate systems to describe and predict the behaviours of visual search. Both models can be mixed together to achieve a better understanding of visual search.

The computational framework shown in Figure 2.11 integrates several current models [2, 67, 108, 113, 114] in order to illustrate the basic and essential components (feature extraction, feature combination, and the attention-selection control of visual features). Further details about these components are given in the following sections.

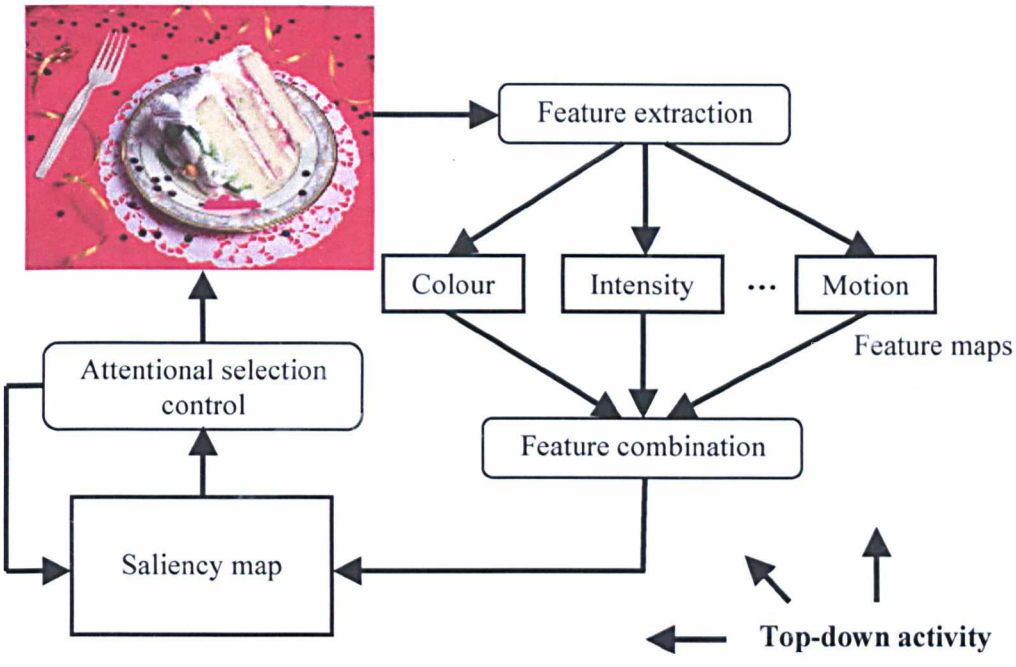



**Figure 2.11** A typical computational model to demonstrate its basic components.

### 2.6.1 Feature Extraction and Feature Maps

Visual search starts with the pre-attentive search of early visual features in a parallel manner. Accordingly, the first processing step in most computational models is to extract visual features, such as colours, intensity, orientations, motion and texture. Some computational algorithms are designed to imitate early visual feature extraction. For instance, the Difference-Of-Gaussian (DOG, Mexican hat) filter [115] is implemented to simulate the response of intensity-centre-surround contrast neurons; and the Gabor-Wavelets filter [116] aims to simulate the response of orientation-selective neurons [108]. In addition to absolute local feature strength, feature contrast is also important, and is associated with the perceptual grouping process. Furthermore, visual search is affected not only by low-level features but also by the high-level semantic properties of objects [117]. In the early stage of visual processing, visual input is decomposed simultaneously into visual features. The results of this decomposition are registered into feature maps. Each feature type has its own feature map. Therefore, information about the spatial distribution and relationship among these features is often neglected in these feature maps.

## 2.6.2 Feature Combination and the Saliency Map

Generally, visual perception can be regarded as a two-fold process. At the first stage, visual stimulus is decomposed into parts that are then integrated back at the second stage [17]. The feature combination and the saliency map (also called the master map or the attention map) reflect the activities at the second stage of visual perception, which is central to most computational models.

The units in the saliency map are arranged topographically in one-to-one correspondence with the units of the retinal image [118]. The value of each unit reflects the spatial competition of visual features, as demonstrated in Figure 2.12. Experimental evidence from physiological studies indicates that the saliency map may exist in the frontal eye field (FEF) region and the posterior parietal cortex [119, 120]. Although recent findings about the distributed representation of top-down and bottom-up information in the brain suggest that the saliency map may be implicitly coded in a distributed network [95], the explicit and centralised saliency map provides an efficient computational strategy for attention-selection control, and is therefore adopted in numerous models [2, 67, 108, 113, 114].

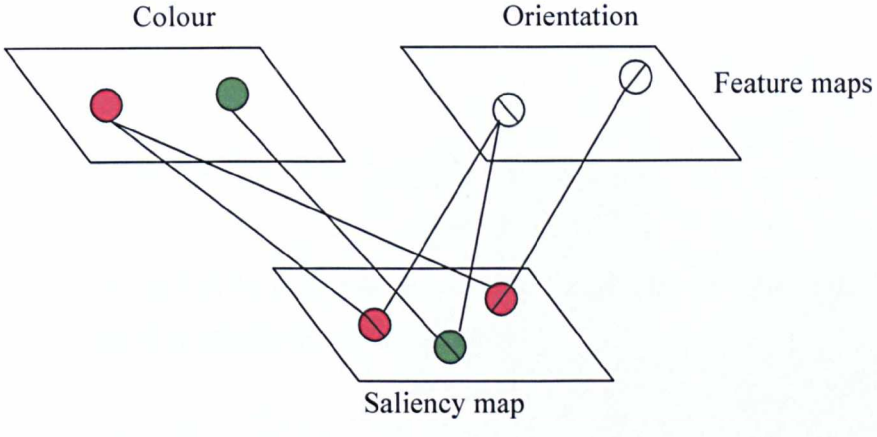



**Figure 2.12** A schematic illustration of the relationship between feature maps and the saliency map. Visual features are independently registered in feature maps. Each unit in the saliency map is arranged in one to one correspondence with the units of the retinal image. The value of each unit in saliency map reflects the spatial competition of visual features, registered in each feature map.

## 2.6.3 Attention-selection Control

### *1) The 'Winner-Take-All' Strategy and the 'Centre-of-Gravity' Effect*

As mentioned above, attention can be regarded as the result of a competition among visual features, and this result can be coded in the saliency map. The 'winner-take-all' strategy directs visual attention towards locations with the highest activity in the saliency map [67, 108]. Therefore, the visual system selects one object in a scene as the focus of attention, and all other objects become part of the background. However, visual search experiments [121, 122] also indicate the 'centre-of-gravity' effect, *i.e.* fixations are often located at the centre of the weighted positions of potential targets. This effect has been taken into account as a supplement to the 'winner-take-all' strategy [123].

Michael *et al.* [118] used connectionist models to analyse spatial attention. Figure 2.13 illustrates the rigid-spotlight attention model used to simulate the 'winner-take-all' competition. In this model, each unit receives an input ($ext_i$) from outside, an excitation ($act_{ii}$) from itself and inhibitions ($act_{ji}$) from other units. At time $t$, the net input ($net_i$) of unit $U_i$ is:

$$net_i(t) = ext_i(t) + \alpha act_{ii} - \beta \sum_{j \neq i} act_{ji} \tag{2.2}$$

The activity update rule is:

$$act_{ii}(t+1) = h(net_i(t)) + (1-\tau) act_{ii}(t),\ 0.0 \leq \tau \leq 1.0 \tag{2.3}$$

where $\alpha$ is the strength of the excitatory self-connection, $\beta$ is the strength of the inhibition between units, and $h$ is a linear threshold function:

$$h(x) = \begin{cases} 0 (x < 0) \\ x (0 \leq x \leq 1) \\ 1 (x > 1) \end{cases} \tag{2.4}$$

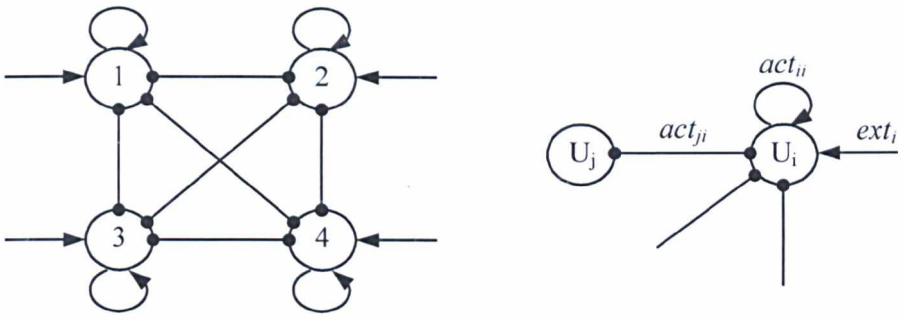



**Figure 2.13** The rigid-spotlight attentional model.

Figure 2.14 shows the activity of the 'winner-take-all' competition between the four units according to this model. The model also works well in simulating attention movement from one place to another when the external inputs of each unit change.

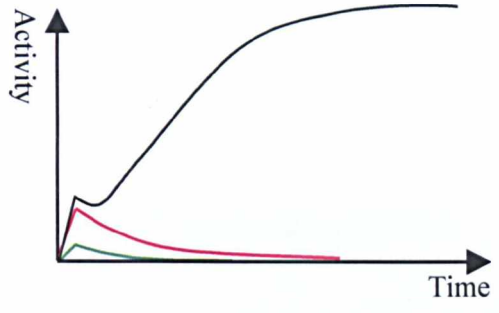



**Figure 2.14** Activity of four attentional units. Unit 1 (black line) has the biggest external input. Unit 3 and unit 4 (two coinciding green lines) have the smallest external input. Unit 2 (red line) has a medium external input.

Instead of using the maximum value in the saliency map to guide attention, FOVEA [13] adopted a probabilistic approach for spatial attention selection. In this model, the 'multiple criteria' map (saliency map) is calculated whenever the next fixation location needs to be determined. The 'multiple criteria' value at each point over the entire visual field is the combination of three criteria: the image-information content criterion, the proximity criterion, and the local suppression criterion. The principle of inhibition of return is implicitly integrated by the local suppression criterion. A random value $x$ is generated by using a modified exponential density $e^{-\alpha(C_{\max}-x)}$, $x \in [0, C_{\max}]$, which guarantees that the value of $x$ is more likely to be in the vicinity of the maximum 'multiple criteria' $C_{\max}$, but is usually not equal to $C_{\max}$. A point with a 'multiple criteria' value $x$ is then selected as the next fixation location according to a proximity-based selection process.

### *2) The efficiency Principle*

According to the efficiency principle, visual attention is deployed to achieve the highest efficiency in terms of search time, minimum eye movements, or maximum information gain. As described in a model proposed recently by Schill *et al.*, attention selection is necessary to resolve ambiguous identities by moving the eyes (or attention) to places with the maximum information for object recognition [124]. Tai *et al.* [125] also assume in their framework that the visual system should always look for, and fixate at, locations with the maximum amount of information. To reduce search time, a maximum-likelihood attention strategy [12] is designed by '*adapting to the distribution of the features in the image and choosing to investigate the feature which is least like the distractor features.*' The efficiency is also related to the coarse-to-fine effect [126-128], which indicates that visual information is processed at different scales, usually starting with coarse (low spatial frequencies, large scale) measurements and ending with finer (high spatial frequencies, small scale) measurements. Rao *et al.* [123] coded this effect into their model and obtained good results in the simulation of saccadic targeting.

### *3) The Constraint of the Short-Term Working Memory*

The limitation of the short-term working memory places a constraint on our visual system. Generally speaking, large memory is usually necessary to achieve an optimal performance. In the case of visual search, previously visited locations ought to be remembered to reduce search time. It has been reported that our visual system uses various scanning strategies to compensate for the limitation of the short-term memory during visual search [107], although those strategies may be inefficient in terms of optimal search time and accuracy.

### *4) Inhibition of Return*

Inhibition of return is a computationally important principle that prevents visual attention from attending eternally to the location with the maximum activity in the saliency map [67, 108, 113, 114] and inhibits the focus of attention shifting back to a location that has just been attended. The inhibition of return needs memory. Because of the limitation of the working memory, the inhibition of return should be treated as a short–term effect in computational modelling, *i.e.* only a few recently visited locations can be memorised.

The control factors listed above are also involved in the updating of the saliency map. This makes the saliency map change over time. The principle of inhibition of return reduces the activity of later-attended locations in the saliency map, and visual attention enhances the saliency of some specific features at the current location in order to resolve ambiguous identities efficiently, thereby resulting in patterns of visual search scan-paths [6, 129].

### 2.6.4 Top-Down Activity

Top-down activity can affect many aspects of a computational model. Each component of the model in Figure 2.11 should take top-down information into consideration. Wolfe [67] uses an activation map to integrate the bottom-up and top-down activations in his Guided Search Model. As in Feature Integration Theory [2], early vision (or the pre-attentive process) registers each feature in the context into its corresponding feature map. The integration of the two activations involves the following calculations.

Bottom-up (stimulus-driven) activation measures how unusual each feature is in the context according to its difference and distance from the other features (the best way of measuring the difference is still under investigation). This tends to direct attention to distinctive items in view, *i.e.*

$$BottomUp = \frac{\sum_{i}^{neighbours} \left( \frac{Diff \times Strength}{Dist} \right)}{\#neighbours} \tag{2.5}$$

On the other hand, top-down (user-driven) activation assigns each feature a weight according to prior knowledge, expectation and experience. Features of the target are given a positive weight. The role of top-down activation in the formation of the activation map is to modulate the bottom-up saliency.

$$Activation = \sum_{i=1}^{features} BottomUp_i \times TopDown_i \tag{2.6}$$

Visual attention will be guided to locations in an order of decreasing activation.

### 2.6.5 System Validation

The objective validation of a computational model is difficult since visual activities such as feature saliency and attention selection are mainly cognitive activities. Current neuroscience techniques are not able to provide a quantitative and objective reference for system validation. However, this can be achieved indirectly by comparing the performance (*e.g.* RT and accuracy) and behaviours (*e.g.* patterns of fixation clusters and fixation sequences) of a computational model with those resulting from human visual search. Stark *et al.* suggests that the first-order Markov matrices and string-editing cost could be used to model and compare fixation sequences [6]. In their recent work [11], they also designed a scheme to evaluate and select bottom-up, context-free algorithms by comparing aROIs (algorithmically detected regions of interest) with hROIs (human identified regions of interest). The methods involved in the algorithm assessment can be employed here for system validation.

### 2.6.6 Signal Detection Theory

Signal detection theory (SDT) [130-132] is another quantitative approach to analysing visual search behaviours. In this theory, internal responses to stimuli (targets and distractors) are regarded as noisy. The discrimination, which depends on the variability and separation of distributions of the noisy responses, is the key to the prediction and explanation of search accuracy, response time, and attention mechanisms. High threshold theory and low threshold theory [133] are applications of SDT to predict visual search behaviours. The high threshold theory can be applied to explain simple visual search tasks, in which basic visual features such as colour and orientation are so easy to detect (they 'pop-out') that distractors can never be mixed up with the target to produce a false alarm; That is, neuron responses to distractors are so weak that they cannot get through the 'high threshold'. '*False alarms arise in behaviour only as a consequence of guessing.*'[133] On the other hand, in the low threshold theory, the responses produced by distractors can be strong enough to pass the 'low threshold', and therefore result in false alarms.

## 2.7 Summary

In this chapter, we have reviewed some of the basic techniques involved in the study of visual search. The dichotomy of the dorsal (where) and ventral (what) streams is essential for understanding the mechanisms of visual processing in the brain. Visual attention separates visual search into pre-attentive and attentive stages. It is guided by both bottom-up and top-down information, and can occur at both early and late vision stages. The units of visual processing can be locations, visual features or objects. Overt eye movement and covert visual attention can be different, but are strongly correlated. Throughout visual processing, attention plays an important role in resource allocation, top-down control, information enhancement, and information binding. As a consequence, almost all computational models concentrate on the modelling of visual attention. To simulate what happens in the brain, most biologically plausible computational models decompose the visual scene into different parts by extracting visual features, and then combine them back into a saliency map. The intrinsic configuration of these models is central to attention-selection control. The brain is capable of adjusting the configuration of its visual system to adapt to different environments according to the demands of search tasks. The understanding of this configuration is essential to both the computational modelling of visual attention [118] and knowledge gathering from human visual search. In specific applications, such as medical imaging, a growing body of research [4] seeks to build intelligent computational vision systems for knowledge-gathering. Understanding the behaviour of human vision system is an important trend in the design of the next generation of computer vision systems.

# Chapter 3

# Scan Path Analysis and Feature Extraction

In the previous chapter, a comprehensive review of visual search methods from a computational perspective has been presented. We have demonstrated that the first stage of visual processing is feature extraction. In studies of visual search, simple and complex cells in the primary visual cortex of the brain have been found to be sensitive to orientation and motion [16, 18, 19] and the functions of these low-level feature detectors have been broadly investigated. In 1987, Young [134] proposed a computational model based on the Gaussian functions of various widths and their partial derivatives to represent the spatial receptive fields in the retina. These functions, like Gabor filters [116], are spatially local and consist of alternating regions of excitation and inhibition of a decaying envelope. From an image-processing point of view, the LOI (Locally Orderless Image) technique [135] has been proposed for image understanding by manipulating the local intensity histogram constructed according to scale-space principles [128]. In medical imaging, a wide variety of feature extractors, ranging from early vision techniques to sophisticated methods, have also been considered for achieving a comprehensive assessment of medical images. In CT lung image understanding, for instance, both the textural appearances and the density of the lung parenchyma have been widely used for diagnosing diffuse lung diseases [136].

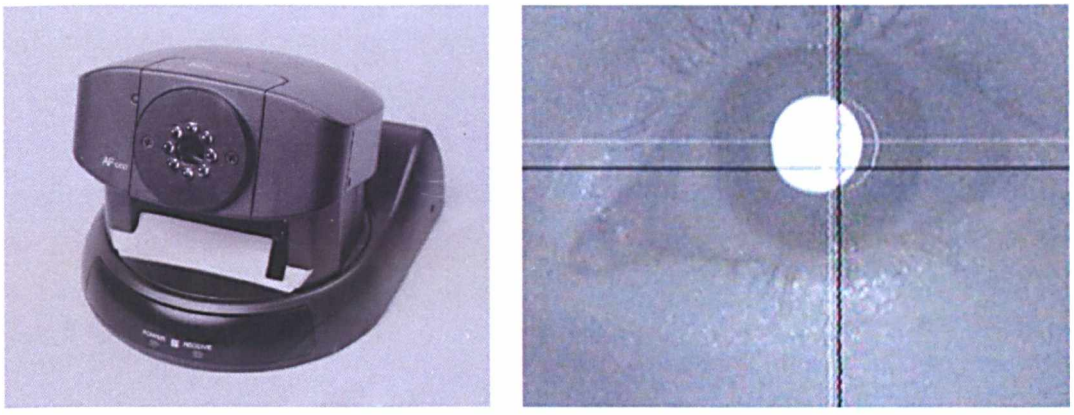
In this chapter, we will focus on fixation analysis and texture feature extraction for characterising the eye movements and image contents of the lung parenchyma on HRCT

images [136]. These are two essential steps in the detection of visual search strategies as described in Chapter 1. The proposed method involves an automated segmentation of the major anatomical structures, including the chest wall, lung parenchyma and macroscopic pulmonary vessels. A DICOM image viewer, implemented for recreating a natural reporting environment, will be presented. The integration of an unobtrusive eye tracking system to acquire reliable scan paths from radiologists is essential to the extraction of visual search strategies.

## 3.1 Eye Tracking and Fixation Analysis

### 3.1.1 Eye Tracking

For all experiments reported in this thesis, the images were displayed sequentially in a 24-inch computer monitor (NEC-MultiSync-FP1370) with a resolution of 1280x1024 pixels. The effective size of the display was 39.4 cm in width and 27.6 cm in height. The observer's eye was at a distance of 65 cm. An ASL Model 504 remote eye-tracking system (Applied Science Laboratories, MA; Sampling rate: 50Hz; Accuracy: 0.5 degrees of visual angle; Resolution: 0.25 degrees of visual angle; Tracking method: Purkinje reflection) has been used to acquire eye gaze coordinates. The Model 504 eye-tracking system is designed to accurately measure the observer's pupil diameter and to track eye movements on a stationary scene [137]. An infrared camera, shown in Figure 3.1, is used to capture the eye image. The pupil is distinguished from the background and the corneal reflection (CR) on the basis of brightness. The separation between the pupil and the corneal reflection varies with eye rotation, but does not vary significantly with eye translation (head movement with respect to the infrared eye camera). Thus, the pupil-CR separation can be used to calculate the direction of the eye gaze, which can be further transformed into the gaze coordinate on the monitor after a calibration procedure. To minimise the amount of head movements, and therefore the degradation of calibration accuracy, a chin rest was used in all the experiments reported in this thesis to ensure an appropriate comfort for the observers.



(a) (b)

**Figure 3.1** The ASL Model 504 remote eye-tracking system. (a) The infrared camera used to capture the eye image. The white circle on (b) indicates the pupil. The white and black crosses indicate the centre of the pupil and the corneal reflection respectively. The direction of eye gaze is calculated from the relative positions of these reflections.

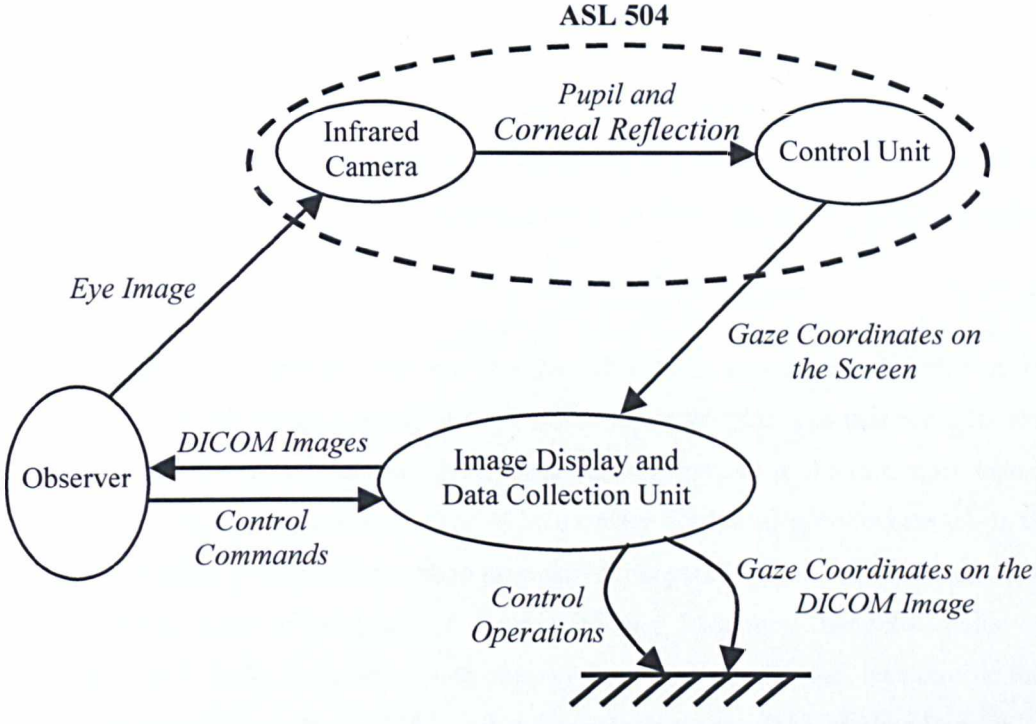



**Figure 3.2** A schematic diagram of the DICOM image-viewing platform.

In order to obtain scan paths when an observer is looking through multi-slice or volumetric CT acquisitions, a DICOM image viewing platform has been implemented. It can simultaneously record both the observer's eye-gaze coordinates and the observer's control operations. This platform has been successfully used to record the scan path on multi-slice CT cases. Figure 3.2 shows its functional diagram. In this platform, the ASL

infrared camera is employed to acquire the pupil and corneal reflections, which are then analysed by the ASL control unit to calculate gaze coordinates on the monitor. The image display and data collection unit simultaneously records both gaze coordinates and control commands by the observer. The recorded gaze coordinates and control commands are synchronised and saved for further analysis.

### 3.1.2 Fixation Analysis

The term fixation denotes an observer's point-of-regard as a stationary object in the visual field is dwelt on by his eyes. Fixations are primarily distinguished from saccades, which are rapid voluntary eye movements from one fixation to another. On the other hand, small involuntary fixational eye movements, including micro-saccade, drift and tremor which occur during a fixation, are difficult to capture without special instruments [138-140]. During the presentation of static scenes, other types of ocular behaviour such as smooth pursuit are rare [32]. The scan path is a sequence of fixations and saccades. This sequence is attributed to the distribution of the photoreceptors on the retina. Since the photoreceptors are packed densely at the fovea, the best visual acuity is only within a visual angle of one to two degrees. In order to build a detailed representation of the image, the eye needs to scan the scene with a series of fixations, resulting in a pattern of scan paths [6].

From a computational perspective, the position of a fixation may be regarded as the centre of a cluster of eye gaze points over a certain period of time. The minimum fixation duration and the maximum fixation boundary are closely related to the minimum latency of saccades and the maximum amplitude of involuntary fixational eye movements. In the literature, a number of criteria have been proposed for separating fixations from saccades. Existing studies have investigated the amplitude and frequency associated with eye movements [139, 140]. In these studies, the eye movement is divided into two distinct groups including the major saccades and the minor saccades. The major saccades are generally larger than 1.2 degrees in amplitude and less than 5 Hz in frequency. The minor saccades are further subdivided into the flicks that are generally between 0.033 degrees and 1.2 degrees in amplitude and the even smaller tremors with a typical amplitude of 0.001 degrees. The flicks and tremors are generally associated with the twitching of the oculomotor muscles. These muscles respond to much higher frequencies, in the range of 40 to 150 Hz.

For this study, a fixation calculation algorithm, which is based on the algorithm proposed in the ASL EYENAL fixation program [137], was implemented. The algorithm uses three boundary thresholds. The first threshold (*Th*1) is applied for starting a fixation when the standard deviation of the visual angles of five consecutive data samples (100 milliseconds) is within it. Once a fixation is initiated, the algorithm uses the second threshold (*Th*2) to terminate the fixation if three consecutive data samples do not fall within the fixation cluster. Finally, the third threshold (*Th*3) is used for filtering out noisy data samples within a fixation to minimise noise and artefacts. Figure 3.3 illustrates a flow chart of the algorithm. The algorithm has three running states: START, INSIDE and OUTSIDE. Initially, the state is set to START. Once the standard deviation of the visual angles of five consecutive gaze points is tested as less than *Th*1, a new fixation is started and the state is changed from START to INSIDE. In the INSIDE state, new samples are included in the fixation if their *Th*2 tests succeed, *i.e.* the difference between the sample's visual angle and the mean value of the visual angles of the gaze points in the temporary fixation set $S_{fix}$ is less than *Th*2. Otherwise, the state is changed to OUTSIDE. The state can also be changed back from OUTSIDE to INSIDE if the *Th*2 test on a new sample succeeds. This process continues until the *Th*2 tests on three consecutive data samples fail.

Figure 3.4 is an example of the fixation analysis. The eye movements were recorded when an observer was asked to search for a waving hand in the crowd. To demonstrate the effect of the three boundary thresholds, we set them to (0.5, 1.0, 1.5) and (0.75, 1.25, 1.75) degrees respectively in the fixation calculation. It is obvious that the algorithm tends to separate a large eye gaze cluster into several fixations when small visual angles are used to determine the boundary thresholds. Throughout this thesis, the three boundary thresholds are set to (0.5, 1.0, 1.5) degrees, as recommended by most eye tracking experiments[4, 5].

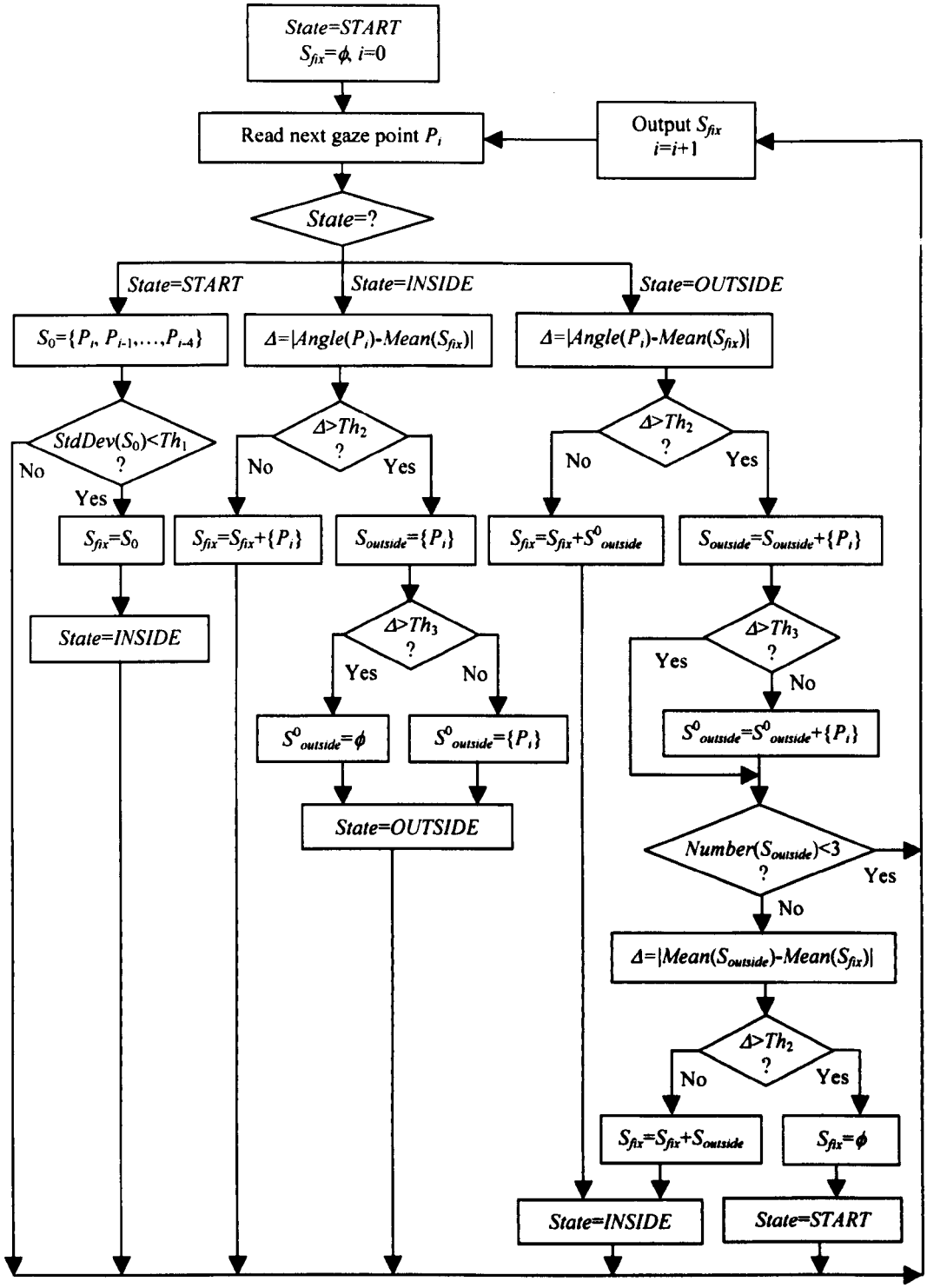



**Figure 3.3** Flow chart of the fixation calculation algorithm. *Th*1, *Th*2 and *Th*3 are three boundary thresholds, which are 0.5, 1.0 and 1.5 degrees respectively. The set $S_{fix}$ stores the gaze points of the current fixation. The set $S^0_{outside}$ is obtained by using *Th*3 to filtering out noise points in $S_{outside}$. The function *Angle*(*P*) calculates the visual angle of the gaze point *P*; *Mean*(*S*) calculates the mean value of the visual angles of the gaze points in the gaze point set *S*; *StdDev*(*S*) calculates the standard deviation of the visual angles of the gaze points in *S*; *Number*(*S*) calculates the number of the gaze points in *S*.

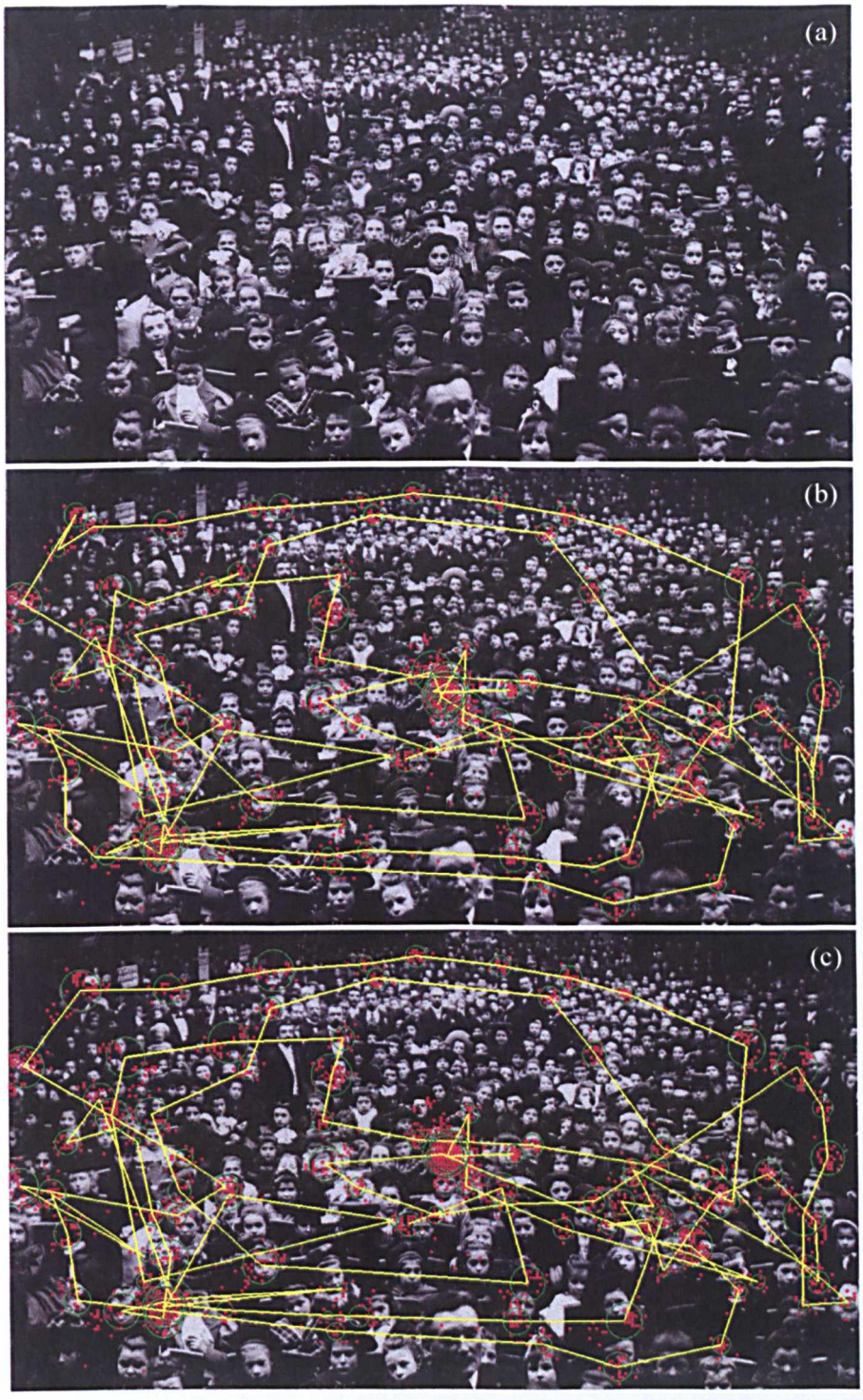



**Figure 3.4** Results of the fixation algorithm shown in Figure 3.3. The eye movements were recorded when an observer was asked to search for a waving hand in picture (a). (b) and (c) illustrate the results of the fixation calculation by setting the three boundary thresholds to (0.5, 1.0, 1.5) and (0.75, 1.25, 1.75) degrees, respectively. The red dots indicate the recorded eye gaze positions; the size of circles represents the dwell time of each fixation and the lines between fixations indicate the saccadic movements. It can be seen that some large fixations in (c) are separated into several adjacent small fixations, indicating the effect of the boundary thresholds.

## 3.2 Image Feature Extraction

For the implementation of the ViTAL framework, feature extraction plays a central role. In general, visual features can be classified as generic features and domain-specific features. The former include features such as colour, shape and texture, while the latter can be application dependent. A great deal of effort has been taken to extract both types of features. In medical imaging, for instance, complicated techniques such as the Half-Max [141] and the ERS (Edge-Radius-Symmetry) transform [142] have also been designed specifically for the identification of the bronchi in CT images. Generic features, especially texture features, have also been widely applied. Landeweerd and Gelsema extracted both first-order features (such as grey level mean) and second-order features (such as grey level co-occurrence matrices) to classify white blood cells [143]. Chappard *et al.* used the run length method and fractal geometry for the reliable detection of bone loss [144]. A study by Chabat *et al.* suggested that the textural characteristics of the lung parenchyma provide essential information for the differential diagnosis of obstructive lung diseases [136, 145]. In this work, they proposed a classification method by using 13 texture features (derived from the histogram, co-occurrence matrix, and run-length parameters) to distinguish between obstructive lung diseases (including centrilobular emphysema, panlobular emphysema, constrictive obliterative bronchiolitis) in thin-section CT images. Due to a large number of feature extractors being published every year, it is difficult to cover all of them in this thesis. We will mainly focus on texture features, which are most relevant to the examples we used in this thesis. It should be noted that the feature selection methods proposed in the following chapters are not designed to favour any particular feature. Domain-specific features and other generic features can also be used without major modification.

### 3.2.1 Texture and Texture-Feature Extraction Techniques

Textures could be regarded as homogeneous visual properties, such as coarseness, directionality and contrast, which we perceive in natural and synthetic scenes. Texture perception plays an important role in human vision to detect and distinguish objects. Although texture analysis has found a wide range of applications and is a basic subject in image processing and computer vision, it has no universally accepted definition [146]. Haralick [147] described the texture of an image according to '*the number and types of its*

*(tonal) primitives and the spatial organisation or layout of its (tonal) primitives.*' Karu *et al.* [146] proposed a local extrema-based method for automatically identifying textured regions. In this method, three types of images (uniform grey value images, white noise images and images containing distinguishable objects) are regarded as non-textured images. The image properties, including *uniformity*, *density*, *coarseness*, *roughness*, *regularity*, *linearity*, *directionality*, *direction*, *frequency* and *phase*, are perceivable to our human visual system and they also play important roles in defining textures [148]. Generally speaking, '*texture is characterised not only by the grey value at a given pixel, but also by the grey value "pattern" in a neighbourhood surrounding the pixel*' [149].

In texture feature extraction, most methods attempt to use values such as mean and variance to quantify image textures. These values can be used to form feature vectors to characterise the underlying structures of a textured region in performing texture segmentation, classification and synthesis [150, 151]. Thus far, a wide variety of extraction methods have been proposed. These methods can be categorised as structural, statistical, model-based and signal-processing methods [151, 152].

- *Structural methods*: The structural models define textures by well-defined primitives, which occur repeatedly according to certain placement rules. In structural texture analysis, the extraction of texture primitives and inference of placement rules are two major steps. In general, this class of models can work only on regular textures, and is therefore most suited to texture synthesis rather than texture analysis. Zucker [153] suggested that natural textures are distortions of ideal textures, which can be transformed to generate the natural texture according to certain placement rules. The process includes two stages: regular mosaic patterns are first generated according to certain rules and then transformed to yield natural textures. Voorhees and Poggio [154] used the Laplacian of Gaussian filter at multiple scales to detect blobs, which are regarded as texture primitives in the image.

- *Statistical methods*: The statistical approaches describe textures according to a set of statistical measures derived from the spatial distribution of grey-level values in the image. The first-order statistical models extract statistics of the individual pixel attributes, while the higher-order models work on the spatial grey level dependences governing the relationships between two or more pixels in the image. As an example, grey-level statistical moments are first-order methods that compute the texture characteristics from the intensity histogram derived from regions of interest, leaving the spatial order of image

pixels as irrelevant. In contrast, high-order texture methods such as the grey-level co-occurrence matrices, which calculate the intensity occurrences of pixel pairs in given directions, rely on the spatial order of image pixels. It has been discovered that humans may be unable to distinguish textures that have the same the first- and second-order statistics but differ in higher-order statistics [155, 156]. Haralick described how textural characteristics can be extracted from the co-occurrence matrices [147, 157]. Other popular statistical methods include the grey level differences [158], grey level run length [159] and autocorrelation [160]. For example, the autocorrelation coefficients of an image $I(x, y)$ is computed as follows:

$$C(p,q)=\frac{MN}{(M-p)(N-q)}\frac{\sum_{i=1}^{M-p}\sum_{j=1}^{N-q}I(i,j)I(i+p,j+q)}{\sum_{i=1}^{M}\sum_{j=1}^{N}I^2(i,j)} \tag{3.1}$$

- *Model-based methods*: Model-based methods use a mathematical model to describe the dependence between a pixel and its neighbourhood. With these methods, the model parameters are used as texture features. The derived mathematical model can be used for both texture description and texture synthesis. For instance, the Markov Random Fields model is based on the assumption that the grey levels at each pixel depend on the grey levels of its neighbourhood [161] and the random field can follow Gaussian or Gibbs distribution [162, 163]. Texture features are extracted by fitting these models to either the entire image (globe models) or small neighbourhoods of pixels (local models). On the other hand, auto-regressive models assume that the grey level of a pixel is a linear combination of the grey levels of its neighbouring pixels [164]. The coefficients resulting from a coarse texture are similar, while the coefficients will be different for a fine texture. The fractal functions have also drawn a great deal of attention in texture modelling and segmentation. Fractals are mathematical sets with a high degree of geometrical complexity [165]. The fractal dimension is scale invariant and is correlated to the human perceivable roughness of image [166], and therefore it can be directly used as a texture measure. Fractal models have been employed in natural image and document analysis [167, 168].

- *Signal-Processing methods*: Psychophysical studies have demonstrated that the visual cortex in the human brain is capable of performing the analysis of frequency, scale, gradient, orientation and motion information. They can also be extracted by using signal-

processing techniques by filtering or transforming the input image into a spatial, frequency or scale domain. In the Fourier frequency domain, for instance, filters can be designed by using annular rings, wedges or strips in different orientations to generate symbolic descriptions of coarseness, homogeneity and directionality [169, 170]. Wavelet and Gabor filters are sensitive to frequencies and orientations within different scales [171, 172]. In the spatial domain, on the other hand, gradient information can be extracted by using operators such as the Laplacian and Sobel operators. This concept was generalised by Laws [148], who developed a set of 'texture energy measures' to determine a variety of features. Below are five 1-D Laws' masks that have been used:

*Level*: [1, 4, 6, 4, 1]
*Edge*: [-1, -2, 0, 2, 1]
*Spot*: [-1, 0, 2, 0, -1]
*Wave*: [-1, 2, 0, -2, 1]
*Ripple*: [1, -4, 6, -4, 1]

In practice, the scale problem imposes a significant challenge to texture analysis as image contents are related to the scale level of observation. A fine or a coarse texture can change its scale if it is observed with increased or decreased resolution. As a result, texture analysis should take image scale into consideration [128, 150]. Texture feature selection is another important issue associated with texture feature extraction. A performance comparison conducted by Weszka *et al.* indicated that co-occurrence matrices outperformed Fourier features for texture classification [158]. This result was also confirmed by Ohanian and Dubes [173], who discovered that the co-occurrence features are generally better for image segmentation than other features, followed by fractal features. They also demonstrated that higher classification accuracy can be achieved by using both co-occurrence and fractal features.

### 3.2.2 The Segmentation of Anatomical Structures

As outlined in Chapter 1, we will use CT lung image analysis as an example for demonstrating the proposed ViTAL framework. The reason we use CT lung image is due to its subtlety and complexity of the image features, and it is a significant challenge in training radiologists. The lungs are a pair of asymmetric, cone-shaped and elastic organs, located in the pleural cavities. Anatomically, the main anatomic structures within the lungs are the pulmonary blood vessels, lung parenchyma and the airways. The

parenchyma refers to the lung tissue as distinguished from the supportive framework. One of the main functions of the lungs is to transport oxygen into the blood and allow carbon dioxide to leave the blood. This process is called gas exchange, which occurs through a process of molecular diffusion across an alveolar membrane. The lungs extend from above the first rib to the superior surface of the diaphragm. The trachea divides into the left and right primary bronchi within the mediastinum. The lungs are separated by the fissures into lobes, which are further divided by incomplete fibrous septa. The right lung has three lobes: right upper lobe (RUL), right middle lobe (RML) and right lower lobe (RLL); while there are only two lobes on the left: left upper and left lower lobes (Figure 3.5). Normal lung parenchyma on a CT scan varies in its density, which is affected by a linear density gradient due to gravity and the phase of respiration. A cardinal CT finding of obstructive lung diseases is the presence of areas of hypo-attenuation of the lung parenchyma. In the case of emphysema, for instance, the hypo-attenuated areas are a direct consequence of lung destruction: the lung parenchyma is replaced by air, leading to a decreased density (or black lung).

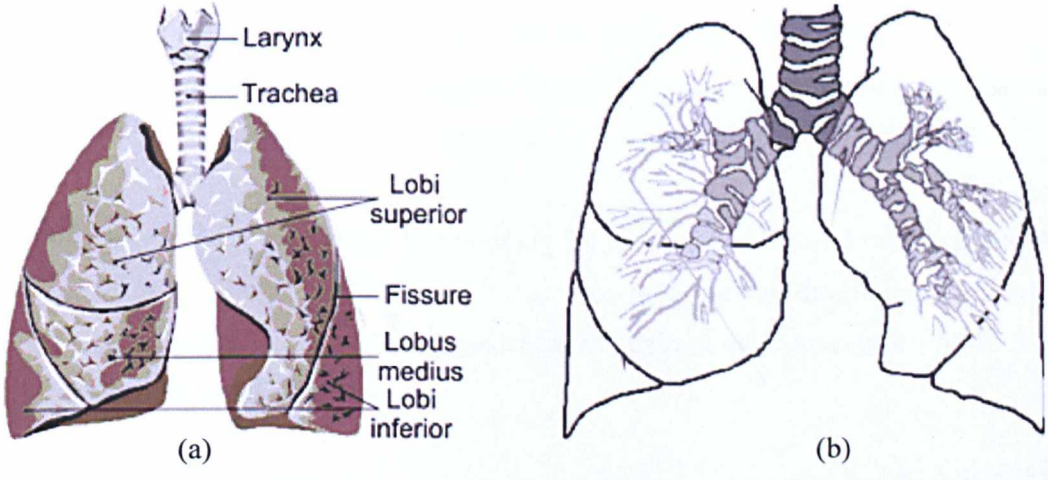



**Figure 3.5** (a) A schematic illustration of the anatomy of the lungs (including the trachea). (b) The bronchial tree (*Adapted from Corel Gallery™ Magic 65,000*).

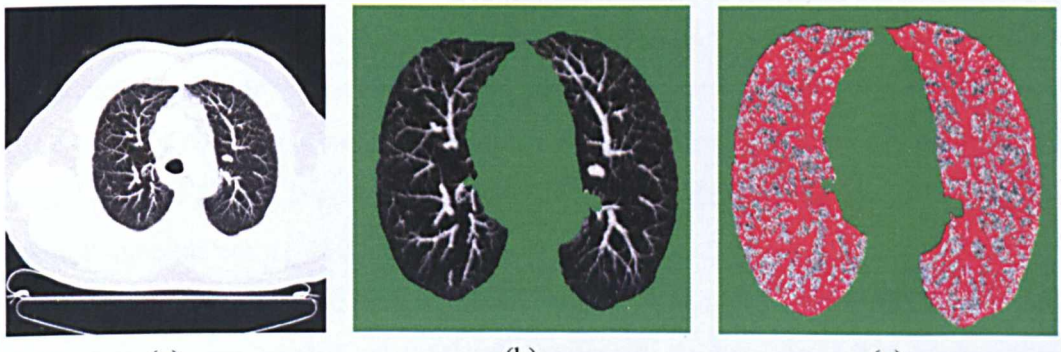
Prior to extracting texture features, the chest wall and pulmonary vessels inside the lung parenchyma need to be removed [5, 136] since they provide no information about parenchyma texture. Therefore, automated segmentation of the major anatomical structures, including the chest wall, lung parenchyma and macroscopic pulmonary vessels, is necessary. Since the CT scanner is accurately calibrated with different tissue densities, the segmentation method to be used is relatively simple and can be based on region-based segmentation with well-defined threshold values.

The density of both the chest wall and the blood vessels is much higher than that of the parenchyma. In a lung HRCT image, the CT value of the chest wall is within the interval [−150, +150]H.U.; the CT value of the blood vessels is within the interval [0, +170]H.U.; and the CT value of the parenchyma is within the interval [−900, −750]H.U. [136]. Because of the large differences between the densities, the thresholding method is able to segment the chest wall and blood vessels from the parenchyma. The threshold values have been empirically set to $T_{chest}$=−300H.U., $T_{vessel}$=−700H.U. in the segmentation algorithm.

The algorithm for segmenting the chest wall from the parenchyma consists of the following four steps:

- Use $T_{chest}$ to perform threosholding operation.
- Apply morphological dilation and erosion operations to eradicate noise and artefacts.
- Explore the image to identify the largest connective white region (region with a CT value larger than $T_{chest}$), which is then regarded as the chest wall.
- Separate the left lung and right lung by calculating the geometric centres of the two largest black regions within the chest wall.

After the parenchyma is extracted by removing the chest wall, the blood vessels inside the lung parenchyma are detected by using $T_{vessel}$ to perform another thresholding operation within the parenchyma. Figure 3.6 illustrates an example of the segmentation of the lung parenchyma and blood vessels.



(a) (b) (c)

**Figure 3.6** Segmentation of the lung parenchyma and blood vessels. (a) original HRCT slice (b) the extracted parenchyma (c) the blood vessels marked with red colour.

## 3.2.3 Feature Extraction

### 3.2.3.1 Texture Feature Extraction

By following the above scheme, the texture features for HRCT lung images have been computed on the segmented images after removal of the chest wall and the macroscopic pulmonary vessels. Sixteen texture features, as listed in Table 3.1, are extracted at every pixel on a HRCT image to capture image textural information including *uniformity, density, coarseness, roughness, regularity, intensity* and *directionality*. The sixteen texture features are calculated at the neighbourhood of each pixel. The radius of the neighbourhood has been empirically set to 22 pixels (corresponding to half a degree of visual angle) and the CT values are quantised into $p$=16 discrete greyscale levels [5, 136].

**Table 3.1** A summary of the 16 texture features used for analysing the HRCT images of the lung.

| Group A | Group B | Group C | Group D |
|---|---|---|---|
| $n^{th}$ order statistical moments | Mean values extracted from 20 co-occurrence matrices | Run-length parameters | Others |
| • mean (mom-mean) | • energy (co-energy) | • short primitive emphasis (run-spe) | • histogram entropy (entropy) |
| • standard deviation (mom-standard-deviation) | • entropy (co-entropy) | • long primitive emphasis (run-lpe) | • fractal dimension (fractal) |
| • absolute deviation (mom-absolute-deviation) | • maximum (co-maximum) | • grey-level uniformity (run-glu) | |
| • skewness (mom-skewness) | • contrast (co-contrast) | • primitive length uniformity (run-plu) | |
| • kurtosis (mom-kurtosis) | • homogeneity (co-homogeneity) | | |

The $n^{th}$ order statistical moments are five descriptive statistical measurements of greyscale-level distribution within each neighbourhood. Other features that describe spatial dependence of greyscale distributions are derived from the set of co-occurrence matrices [147, 157]. Co-occurrence matrices are based on the estimation of the probability of going from one grey level to another grey level given specific inter-sample spacing and direction. A set of 20 co-occurrence matrices $\{C_{1\le i\le 4,\ 1\le j\le 5}\}$ along four directions from $i$=1 to $i$=4, representing 0°, 45°, 90° and 135° respectively, and five distances ranging, from $j$ =1 to $j$ =5 pixels, have been calculated. For each co-occurrence matrix, five scalar measurements, namely *energy*, *entropy*, *maximum*, *contrast* and *homogeneity*, are extracted to reduce the dimensionality of the textural features according to the following equations:

$$energy(C) = \sum_{i,j} C_{i,j}^2 \tag{3.2.a}$$

$$entropy(C) = \sum_{i,j} C_{i,j} \log C_{i,j} \tag{3.2.b}$$

$$maximum(C) = \max_{i,j} C_{i,j} \tag{3.2.c}$$

$$contrast(C) = \sum_{i,j} |i-j|^{\alpha} C_{i,j}^{\beta}, \text{ with } \alpha=1, \beta=1. \tag{3.2.d}$$

$$homogeneity(C) = \sum_{i,j} \frac{C_{i,j}}{1+|i-j|} \tag{3.2.e}$$

The five mean values are the averages of these scalar measurements over 20 co-occurrence matrices.

Since co-occurrence matrices do not capture the shape information of the grey-level primitives (*i.e.* a maximum contiguous set of constant grey level pixels located in a line), grey-level run-length features are also computed to measure the properties of linearly adjacent pixels having the same grey level [159]. In run-length calculation, consecutive pixels of the same grey value in a given direction constitute a run. The run-length parameters used in this work include *short primitive emphasis* (*spe*), *long primitive emphasis* (*lpe*), *grey-level uniformity* (*glu*) and *primitive length uniformity* (*plu*), which are calculated according the following equations [174]:

$$spe = \frac{1}{B_{tot}} \sum_{a=1}^{p} \sum_{r=1}^{r_{max}} \frac{B(a,r)}{r^2} \tag{3.3.a}$$

$$lpe = \frac{1}{B_{tot}} \sum_{a=1}^{p} \sum_{r=1}^{r_{max}} B(a,r) r^2 \tag{3.3.b}$$

$$glu = \frac{1}{B_{tot}} \sum_{a=1}^{p} \left( \sum_{r=1}^{r_{max}} B(a,r) \right)^2 \tag{3.3.c}$$

$$plu = \frac{1}{B_{tot}} \sum_{r=1}^{r_{max}} \left( \sum_{a=1}^{p} B(a,r) \right)^2 \tag{3.3.d}$$

where $B(a,r)$ is the total number of run-length primitives with the grey-level $a$ and length $r$; $r_{max}$ denotes the maximum primitive length and $B_{tot}$ is the total number of runs:

$$B_{tot} = \sum_{a=1}^{p} \sum_{r=1}^{r_{max}} B(a,r) \tag{3.4}$$

*Short primitive emphasis* (*spe*) measures the predominance of short runs in a textural pattern; a large *spe* indicates a fine textural pattern or a high proportion of short runs. *Long primitive emphasis* (*lpe*) measures the predominance of long runs in a textural pattern; a large *lpe* indicates a coarse textural pattern. The *uniformity* features (*glu* and *plu*) measure the similarity of the grey-level values and the run lengths in a textural pattern. A small *glu* or *plu* value denotes that the grey-level values or the run lengths are similar in the pattern.

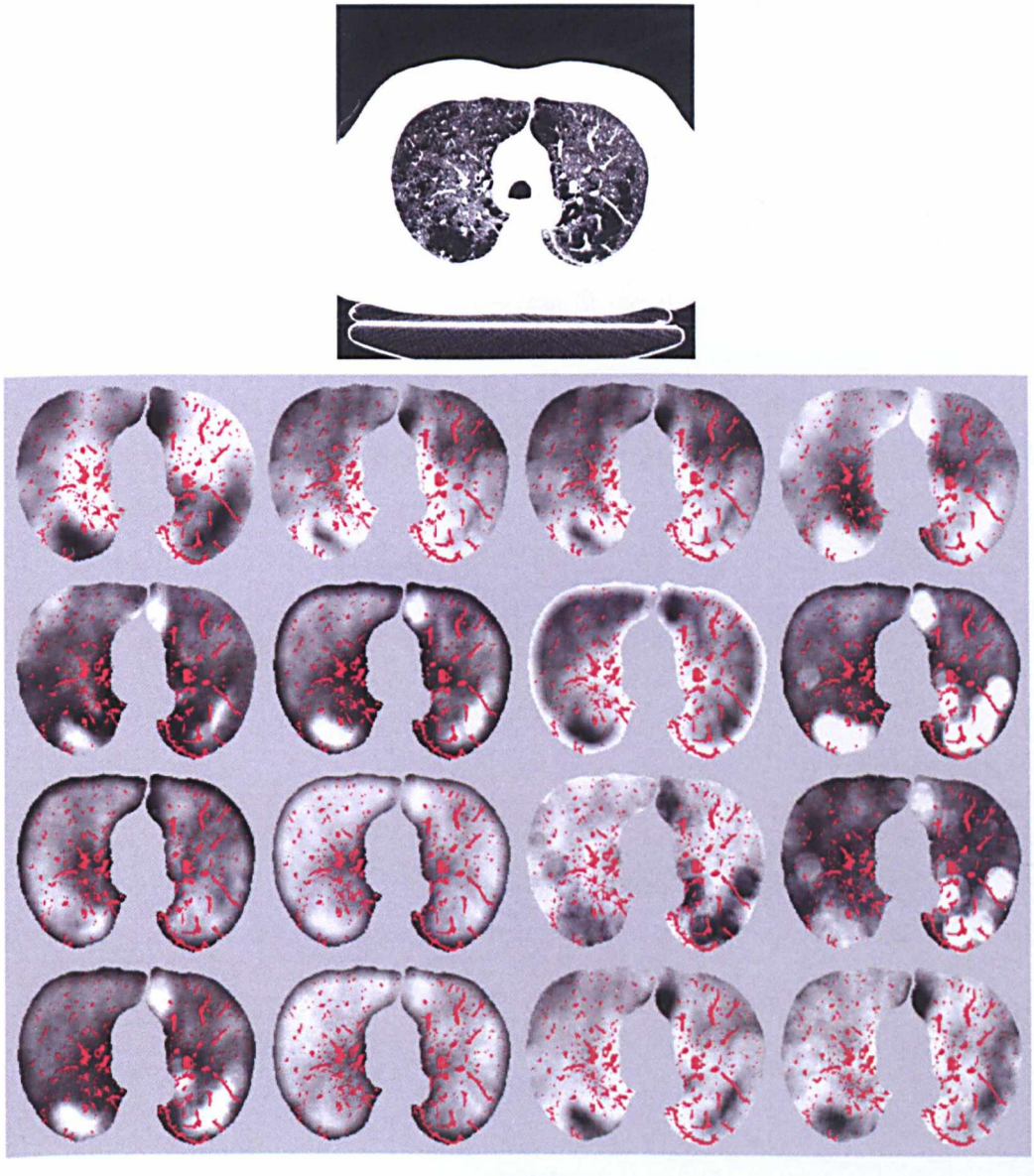


**Figure 3.7** A CT image and its extracted texture features by using the sixteen feature extractors summarised in Table 3.1.

In order to capture the *roughness* information of the neighbour region around each pixel, the feature library also includes fractal dimension, which is calculated by using the blanket technique [167], and the entropy of the local greyscale-level histogram are also derived. Figure 3.7 illustrates a CT image and its texture appearance by applying the 16 texture extractors.

### 3.2.3.2 'Texture of Texture' Feature Extraction

One of the major concerns of the feature extraction library in the ViTAL framework is the completeness of the feature extraction. The feature extractors mentioned above are determined by domain experience, whereas in other situations this may not be optimal. Since an exhaustive list of feature extractors is not possible, we have considered a generic multi-scale feature extractor, called 'texture of texture' in this thesis. The purpose of introducing the 'texture of texture' features, which were originally proposed by DeBonet *et al.* for image retrieval tasks [14], is to construct a large image feature library for generic purposes in a purely bottom-up fashion. It has been reported that these features are selective to the 'sparse' causal structures of images and have a strong similarity with the response properties of the visual cortex [175]. They are calculated in an iterative way through a tree of filtering operations in a filter network responding to certain textural structures in the image. The feature generation process is based on a set of 'first-order' filters shown in Figure 3.8, including 'oriented edges', 'centre surround' and 'bars', *i.e.*

$$[1/4 \quad 2/4 \quad 1/4], [1/2 \quad -1/2], \begin{bmatrix} 0 & 1/4 & 0 \\ 1/4 & 0 & -1/4 \\ 0 & -1/4 & 0 \end{bmatrix}, [1/4 \quad -2/4 \quad 1/4], \begin{bmatrix} 0 & 1/4 & 0 \\ -1/4 & 0 & 1/4 \\ 0 & -1/4 & 0 \end{bmatrix}$$

These first-order filters are obtained by combining the above five filters horizontally and vertically in all possible ways. For instance, by combining $[1/4 \quad 2/4 \quad 1/4]$ with $\begin{bmatrix} 0 & 1/4 & 0 \\ -1/4 & 0 & 1/4 \\ 0 & -1/4 & 0 \end{bmatrix}^T$, we have

$$\frac{1}{16}\begin{bmatrix} 0 & -1 & -2 & -1 & 0 \\ 1 & 2 & 0 & -2 & -1 \\ 0 & 1 & 2 & 1 & 0 \end{bmatrix}$$

This forms a kernel displayed at the top-right corner of Figure 3.8.

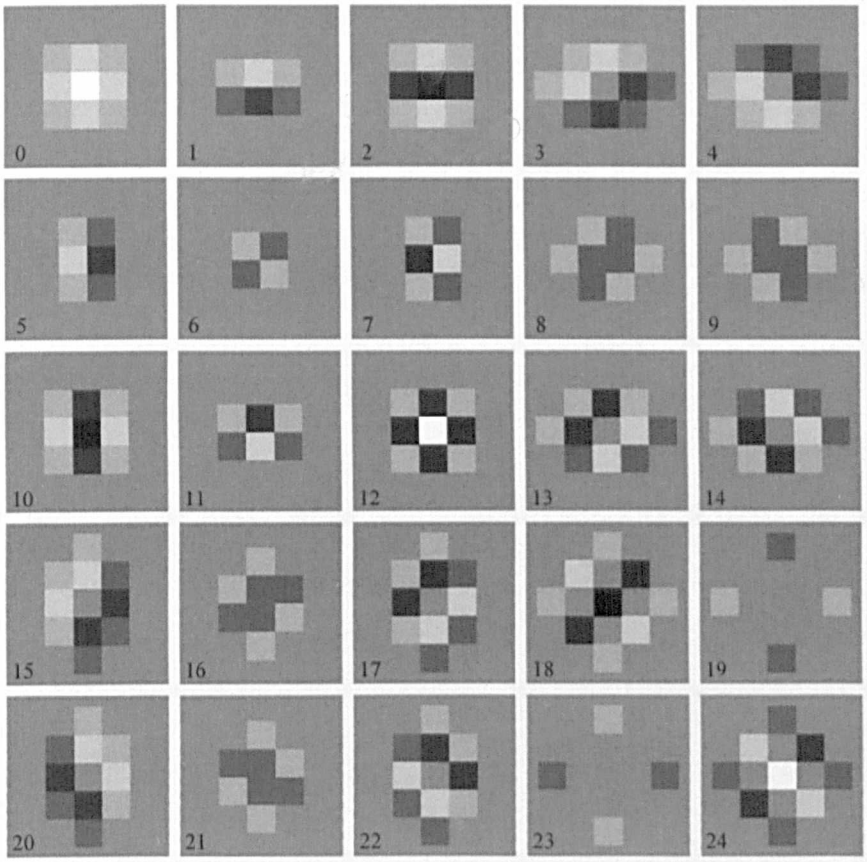



**Figure 3.8** The 25 'first order' filters used in computing the feature maps, numbered from 0 to 24.

The computation of 'texture-of-texture' features is straightforward. At the highest level of resolution, the image is first convolved with the 25 'first-order' linear filters and then followed by a down-sample operation, resulting in 25 high-scale feature maps. Each of them characterises the local textural structure of the input image to be used as the input for the next round of feature extraction. By repeating convolution and down-sampling on each of these 25 low-resolution feature maps, a total of 625 higher scale 'texture-of-texture' feature maps are produced in the second layer. The idea behind this recursive process is that each layer 'discovers' the spatial arrangements of textures specified in the

previous layer. Formally, these features are calculated from an image by using the following formulae:

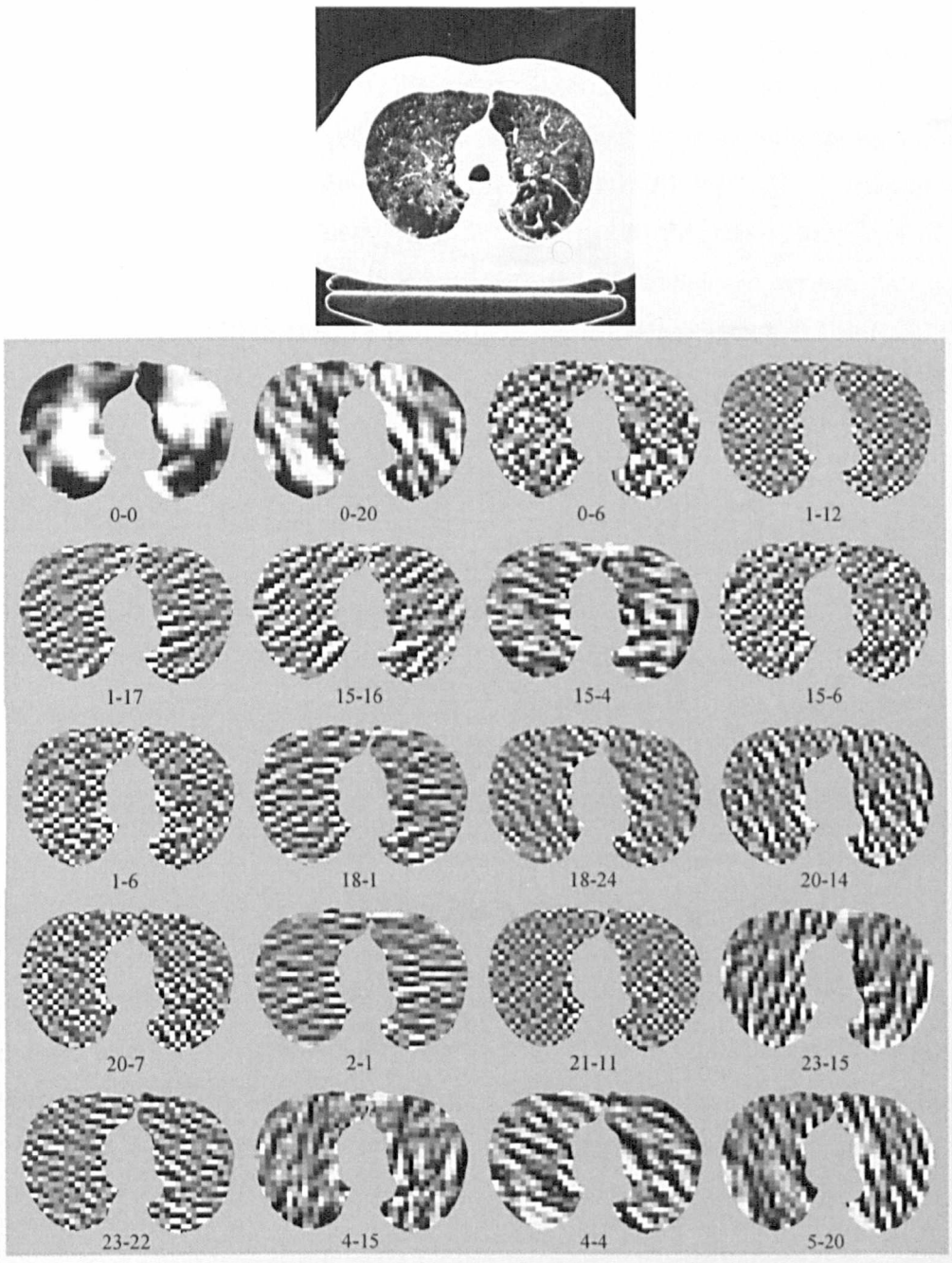



**Figure 3.9** 20 selected 'texture of texture' feature maps of the CT image. The number pairs under each map indicate how the maps are generated. For instance, '23-15' indicates that the map is generated by applying filter 23 and filter 15 (shown in Figure 3.8) sequentially.

$$\begin{cases} M_{i,j,k} = \downarrow \left( \left| f_k \otimes M_{i,j} \right| \right) \\ M_{i,j} = \downarrow \left( \left| f_j \otimes M_i \right| \right) \\ M_i = \downarrow \left( \left| f_i \otimes I \right| \right) \end{cases} \tag{3.5}$$

where $I$ is the image, $f$ is one of the 'first-order' filter, $\otimes$ is the convolution operation and $\downarrow$ is the down-sample operator. With this framework, feature maps on higher layers carry the arrangement information of the feature maps onto lower layers and the down-sample operations increase the scale level of the feature maps. In this thesis, the factor of the above down-sample operator is 3 pixels in both the horizontal and vertical directions. Figure 3.9 illustrates 20 selected feature maps of the CT image shown in Figure 3.7 (the details of how these 20 feature maps are selected will be presented in Chapter 7).

## 3.3 Conclusions

In this chapter, we have introduced a platform for eye-movement tracking and methods for image-feature extraction. They form the basis of the detection of visual search strategies. Texture features are important visual features that have been found to be sensitive to abnormalities related to obstructive and diffuse lung diseases. A set of sixteen textures is included in the feature library. The 'texture of texture' features are highly selective to the causal structures of the images and are calculated in a recursive way. This approach is thus suitable for building a large and informative feature library.

It should be noted that visual search can be performed at different scales [1, 128]. It is, therefore, also necessary to analyse scan path at different scale levels. The proposed feature extractors have been determined empirically for limited scale levels. However, this may not be sufficient for a complete understanding of visual search strategies. To overcome this problem, scale-space theory should be introduced such that the scale is treated as a continuous variable in the design of feature extractors. It is worth noting that the completeness of the feature library is an important issue to consider. Although hundreds of features can be calculated by using the above methods, we cannot be certain that the image contents are well represented for the analysis of visual search strategies. Therefore, an automatic way to achieve feature construction and detection needs to be investigated [176, 177].

# Chapter 4

# Framework of Hot Spot Detection

As discussed in Chapter 2, what we are looking at and where we should direct our visual attention are two key questions that our visual systems are constantly faced with during visual search tasks. The accuracy and speed of the search depends heavily on the strategy of deploying visual attention in response to visual search tasks. Visual search is a process that takes time, and the nature of the process is governed by our knowledge, interests, and expectations of the scene. It is generally accepted that the process of human visual search can be divided into two stages: pre-attentive search and attentive search. Pre-attentive search is a fast and parallel operation over the entire visual field of view to register primitive visual features such as colour and orientation into the memory. Attentive search operates slowly and serially over small parts of the visual field, dictated by the limited ability of the visual system. The covert visual attention and overt eye movements are different but closely related [1]. Scan paths generated from eye movements usually have repetitive and idiosyncratic patterns [6, 7]. They provide important information about the underlying strategies used for visual search.

As introduced in Chapter 3, a typical scan path consists of a series of saccades and fixations. Saccades are fast and ballistic eye movements from one fixation to another. Information is gathered mainly during fixations and very little information is obtained during saccades. Consequently, most current approaches to visual search research are focused on the analysis of fixation position (saccade landing position), fixation dwell time, and fixation sequences. Computationally, clustering methods can be used to detect regions of interest based on the analysis of fixation landing positions and their associated

dwell time. For example, White *et al.* performed eye tracking for understanding the visual search behaviour of radiologists to detect lookpoint and dwell clusters of scan paths in spatial domain [178, 179]. In a separate study, Stark *et al.* [129] suggested the use of first-order Markov matrices and string-editing theory for modelling fixation sequences. One inherent limitation of using the spatial distribution of scan paths is that in practice it is not always possible to form meaningful fixation clusters. In radiology, for instance, experienced radiologists do not appear to scrutinise an image thoroughly, but rather linger on key areas, leaving large areas of the image unexamined. It has been shown that experts examine fewer fixation points compared to trainees or unskilled observers [8, 9]. After an area of interest or abnormality is fixed upon, there follows a rapid checking of the validity of the observation to exclude technical causes or artefacts and, more importantly, the search for ancillary and corroborative signs to support the initial observation. For many symptoms, the manifestation of the disease can form distributed ancillary signs and the resulting scan paths can have contrastingly different spatial distributions, even when the same strategies are followed.

In this chapter, we propose a framework for identifying salient visual features derived from visual search scan paths. The key difference of the proposed method in comparison to existing approaches is that it is based on the modelling of fixations in feature space, and therefore unlocks the intrinsic link between overt eye movements and visual attention. Similar to the two-stage visual search theory, the proposed method first projects scan path into feature space, which is then followed by identifying 'hot spots' where visual attention has focussed. The results can be projected back into the spatial domain, allowing the identification of salient regions of interest. The main motivation of the study is to improve the quality of visual decision support systems by encapsulating and utilising those factors that are either consciously or subconsciously applied by experienced radiologists during visual assessment.

# 4.1 Material and Methods

## 4.1.1 Representation of Visual Search Scan Path

### *1) Scan Path*

A scan path is a sequence of quadruples $<x_i, y_i, d_i, t_i>$ $(1 \le i \le L)$, where $L$ represents the number of fixations, $x_i$, $y_i$ represent the spatial location of each fixation, $t_i$ represents the corresponding dwell time, and $d_i$ refers to information such as pupil diameter and other related measurements. For simplicity, we will not consider $d_i$ in this study and therefore the scan path is defined as $<x_i, y_i, t_i>$ $(1 \le i \le L)$.

### *2) Features*

The properties of an image can be described by its features, which may be low-level features, such as intensity, colour, and texture, or high-level features that are related to certain anatomical structures in medical images. In general, features can be divided into three categories: interval (the distance between features is known), ordinal (the order between features is known), and nominal (the distance and order between features are unknown). Texture features can be regarded as interval whereas abstract high-level features are usually considered to be nominal. We use $F_1, F_2, \ldots, F_N$ to represent features for a given image extracted by $N$ feature extractors, with $\boldsymbol{f} = (d_1, d_2, \ldots, d_N)$ denoting a feature point in the space formed by $F = F_1 \times F_2 \ldots \times F_N$. Each feature $d_k \in F_k$ $(1 \le k \le N)$ can be either continuous or discrete.

### *3) Feature Space Projection*

Every pixel covered by the scan path can be projected to the feature space by considering its associated feature vector. Given a scan path $\xi = <x_i, y_i, t_i>$ $(1 \le i \le L)$, the projection $<x_i, y_i> \rightarrow \boldsymbol{f}_k$ indicates that the fixation centred at $<x_i, y_i>$ has a feature value of $\boldsymbol{f}_k$. A probability function $P^i_{f_k}(x, y)$ is normally assigned to the projection $<x_i, y_i> \rightarrow \boldsymbol{f}_k$, which denotes the probability of a feature point $\boldsymbol{f}_k$ being targeted at fixation $<x_i, y_i>$. To take into account all pixels falling within the foveal field of $<x_i, y_i>$, the value of $P^i_{f_k}(x, y)$ can be estimated by using the 'hit' method as described by Kundel *et al.* [180], where all pixels within a

1.5° visual angle around $<x_i, y_i>$ are considered. As the visual acuity drops off dramatically from the centre of focus, it is natural to model the foveal field activation as a Gaussian distribution [181]. Features corresponding to pixels falling within the foveal field of view are assigned a non-zero probability with value inversely proportional to the spatial distance to $<x_i, y_i>$, as dictated by the Gaussian function. Given a scan path $\xi$, its total dwell time on a feature point $f_k$ is defined as

$$T(f_k) = \sum_{\langle x_i, y_i \rangle \in \xi} \left( \sum_{\langle x,y \rangle \in \Omega(x_i, y_i)} t_i P^i_{f_k}(x,y) \right) \tag{4.1}$$

where $\Omega(x_i,y_i)$ is the foveal field of fixation $<x_i, y_i>$, and $P^i_{f_k}(x, y)$ is a Gaussian function centred at $<x_i, y_i>$.

### 4.1.2 The Fixation Density Function

As known from visual search theory, the selection of visual attention is biased towards regions sharing common features with the target [30]. The correct identification of this bias is important towards the understanding of the cognitive process involved in visual search. To this end, we introduce a fixation density function ($\Gamma$) defined in feature space $F$ for extracting salient visual features. For this study, we assume that the saliency of a given feature is proportional to the total dwell time defined in (4.1), *i.e.* given a scan path $\xi = <x_i, y_i, t_i>$ $(1 \le i \le L)$ and a feature point $f_k$,

$$\Gamma(f_k) = \alpha_k T(f_k) \tag{4.2}$$

where

$$\Gamma(f_k) \ge 0 \text{ and } \sum_k \Gamma(f_k) = 1 \tag{4.3}$$

In (4.2), $\alpha_k$ is a normalisation factor depending on the context of the scene. The fixation distribution in the feature space can be affected by attention selection during visual search and projection bias. Projection bias is independent of visual search strategies and is only influenced by the relative abundance of different features within the image. This can be

understood by considering an oversimplified case of searching for white flowers among a sea of red roses. When there is no clear search strategy being adopted, nearly all parts of the image will be scanned and the total dwell time on red would be significantly larger than that of white. This does not mean that red is the primary feature leading to the identification of all white flowers. Due to the abundance of the roses, there is a higher probability of fixation points landing on the red colour, thus introducing a projection bias. To analyse visual search behaviour in feature space, projection bias must be eliminated in order to uncover the underlying strategies used. Quantitatively, this bias is related to $T_0(f_k)$ derived from a scan path that covers the entire image through a uniform sampling procedure without feature preference [107], *i.e.*

$$T_0(f_k) = \sum_{\langle x_i, y_i \rangle \in \text{Image}} \left( \sum_{\langle x, y \rangle \in \Omega(x_i, y_i)} t_0 P^i_{f_k}(x, y) \right) \tag{4.4}$$

where $t_0$ is a constant due to the uniform sampling and can be considered as a unit time, *i.e.* $t_0$=1. The use of uniform coverage for deriving $T_0(f)$ means that no part of the image was over/under scanned. Since the order with which each part of the image is visited is not important to the derivation of $T_0(f)$, either random or sequential coverage of the image will provide similar results for $T_0(f)$. Sequential coverage was used in the study presented in this chapter. In this case, there is no specific attention given to individual visual features. The corresponding fixation density will therefore have a uniform distribution, *i.e.*

$$\alpha_k T_0(f_k) = \Gamma_0 \tag{4.5}$$

By substituting (4.5) into (4.2) and using (4.3) to normalise Γ, the scene-specific normalisation factor $\alpha_k$ can be eliminated, leading to

$$\Gamma(f_k) = \frac{T(f_k)/T_0(f_k)}{\sum_k (T(f_k)/T_0(f_k))} \tag{4.6}$$

The inherent meaning of (4.6) can be explained by using the signal detection theory [130]. If $T_0(f_k)$ is viewed as a noise response and $T(f_k)$ as that of signal-plus-noise, the ratio between $T(f_k)$ and $T_0(f_k)$ is the likelihood ratio of the decision theory. Here, 'signal' refers to attention in terms of dwell time on visual features. The essence of signal

detection theory is to detect a signal by comparing the noise response to the distribution that contains both signal and noise. There are several ways to measure the dissimilarity between distribution $T(f_k)$ and $T_0(f_k)$. One method is to use information theory based measures such as the Kullback-Leibler ($KL$) divergence, which is defined as

$$KL(T,T_0)=\sum_k T'(f_k)\log_2\left(\frac{T'(f_k)}{T_0'(f_k)}\right) \tag{4.7}$$

where $T'(f_k)$ and $T'_0(f_k)$ represent normalised $T(f_k)$ and $T_0(f_k)$, respectively. The value of $KL(T, T_0)$ is non negative, and $KL(T, T_0)=0$ if and only if $T=T_0$.

It is also possible to use other measures such as Jensen-Shanon divergence [182], Chi-Square statistics or the area under Receiver Operating Characteristic (ROC) curve, which reflects the reliability of separating these two distributions, or in other words, the intrinsic ability of the features to discriminate scan path driven by a clearly defined strategy from that of a random walk, to measure the dissimilarity between distribution $T(f_k)$ and $T_0(f_k)$.

### 4.1.3 Hot Spot Analysis

#### *1) Feature Extraction and Selection*

With the above definition of the information content carried by the scan path, we can now identify salient features used for visual search. The framework used in this study for decision support is schematically illustrated in Figure 4.1, where a feature library is used for extracting the feature vector for each pixel.

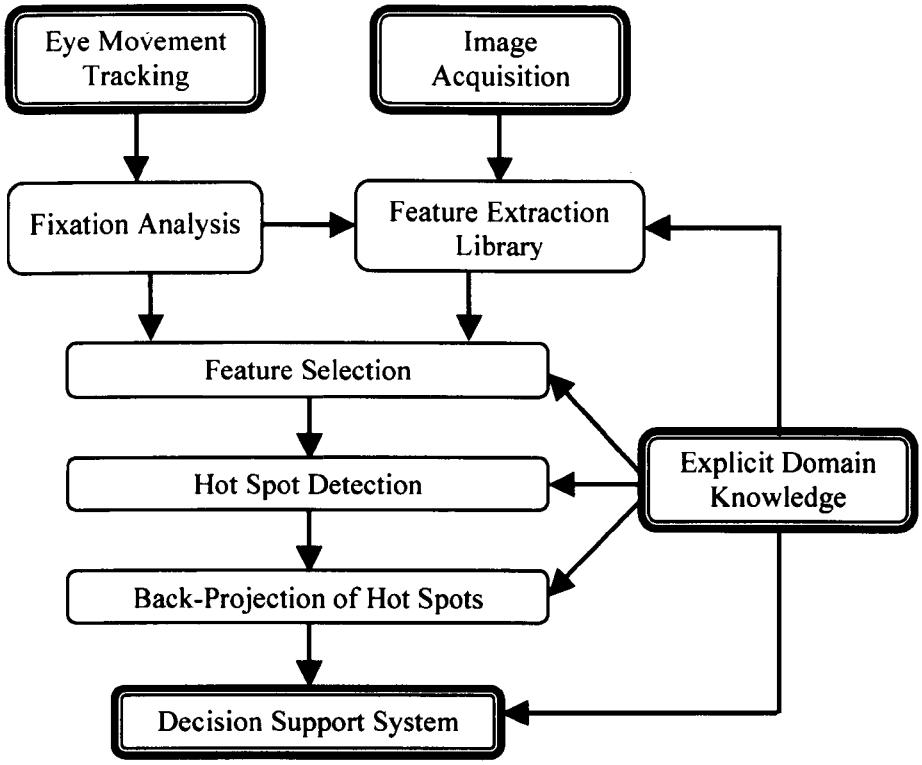



**Figure 4.1** The overall framework of the proposed method for feature and attention selection based on visual search scan path analysis. Explicit domain knowledge for this task was utilised to define texture dimension for the feature space.

After a set of features is extracted, a selection procedure is necessary to determine an appropriate subset of features that are salient. This feature subset defines a feature subspace where visual search is embedded. In some circumstances, the relevant features used in visual search are obvious. For instance, in the example given above the visual feature used is colour. The selected feature subset should contain the least amount of features that can unambiguously identify visual attention. Based on (4.7), the $KL(T, T_0)$ value is determined by the fixation distribution and the context of the background scene. A feature subset with a large $KL(T, T_0)$ value indicates feature preference during visual search. A small $KL(T, T_0)$ value indicates no feature preference or the features in the subset are not sufficient for identifying visual attention. Given $N$ features, the number of candidate feature subsets is $2^N$; therefore, the selection process can be computationally prohibitive. To circumvent this problem, we propose a simple forward feature selection method in this section. The algorithm starts with an empty feature set and at each iteration a new feature is added to the feature subset if the corresponding $KL(T, T_0)$ value is the largest when this feature is combined with the existing feature set. Otherwise, another feature is selected from the candidate feature set. The algorithm repeats until a feature set has a large enough $KL(T, T_0)$ value or when all features have been tested. There is no

guarantee that the algorithm will converge to the optimal feature set and it is also possible that the $KL(T, T_0)$ value of the resultant feature set is not significant. In this case, new features must be incorporated. Such a situation may also be met when there is no clear feature preference during visual search tasks.

### *2) Hot Spot Detection and Back-Projection*

After identifying the feature space used for visual search, Equation (4.6) is used to calculate the fixation density of the selected features. Feature points with a relatively high fixation density value are considered to be important, or 'hot'. According to signal detection theory, the selected threshold should ensure that the likelihood ratio $T(f_k)/T_0(f_k)$ is large enough. Thus, simple thresholding can be utilised for the detection of hot spots, *i.e.* a feature point whose likelihood ratio $T(f_k)/T_0(f_k)>\beta$ can be regarded as salient. The threshold $\beta$ can be determined according to the trade-off between the false-alarm rate and the miss rate in signal detection theory. If the cost of false-alarm and the cost of miss are equal and no prior knowledge about the saliency of features is available, the optimal threshold becomes $\beta$=1.0. Once the hot spots in the feature space are identified, they can be back-projected to the spatial domain for highlighting regions in the original image that contains such features. Domain knowledge from radiologists can be utilised here to explain the diagnostic meaning of each salient feature spot.

### *3) Pattern Integration*

To obtain the common pattern embedded in the scan paths acquired from different observers, we need to integrate search patterns detected from each scan path. This integration is constrained by the common knowledge of the observers and the information of the background.

If a set of scan paths $\xi_m=<x_{im}, y_{im}, t_{im}>$ $(1\le im\le L_m,\ 1\le m\le M)$, collected from a group $(M)$ of observers with similar prior-knowledge and experience, is available, a set of total dwell times on feature $f_k$, $T_m(f_k)$ $(1\le k\le N,\ 1\le m\le M)$, can be calculated for each scan path by using (4.1). The overall total dwell time of this group of observers on feature $f_k$ is defined as the sum of the total dwell time of each observer in a given period of time. During the given period, each observer adopts the same strategy as used when searching individually. Thus, the sum of total dwell time on $f_k$ in a unit time can be estimated as

$$\begin{cases} TS(f_k) = \sum_{m=1}^{M} T_m'(f_k) \\ T_m'(f_k) = \dfrac{T_m(f_k)}{\sum_{f_k} T_m(f_k)} = \dfrac{\sum_{(x_{im},y_{im})\in\xi_m} \left( \sum_{(x,y)\in\Omega(x_{im},y_{im})} t_{im} P_{f_k}^{im}(x,y) \right)}{\sum_{f_k} \left( \sum_{(x_{im},y_{im})\in\xi_m} \left( \sum_{(x,y)\in\Omega(x_{im},y_{im})} t_{im} P_{f_k}^{im}(x,y) \right) \right)} \end{cases} \quad (4.8)$$

By normalising *TS*, we have the integrated total dwell time:

$$TI(f_k) = \frac{\sum_{m=1}^{M} T_m'(f_k)}{\sum_{f_k}\sum_{m=1}^{M} T_m'(f_k)} = \frac{\sum_{m=1}^{M} T_m'(f_k)}{\sum_{m=1}^{M}\sum_{f_k} T_m'(f_k)} \quad (4.9)$$

Since $\sum_{f_k} T_m'(f_k) = 1$, we have

$$TI(f_k) = \frac{1}{M}\sum_{m=1}^{M} T_m'(f_k) \quad (4.10)$$

Thus, the integrated total dwell time is the sum of the total dwell time divided by the number of observers. Accordingly, the likelihood ratio between $TI(f_k)$ and $T_0(f_k)$ is

$$\frac{TI(f_k)}{T_0(f_k)} = \frac{1}{M}\sum_{m=1}^{M} \frac{T_m'(f_k)}{T_0(f_k)} \quad (4.11)$$

This indicates that the integrated likelihood ratio is also the sum of the likelihood ratio divided by the number of observers.

Based on the above derivation, it is now possible to use the integrated likelihood ratio to detect hot spots for pattern integration. Given a threshold value $\beta$, we need to test the value of

$$\frac{TI(f_k)}{\beta T_0(f_k)} = \frac{1}{M}\sum_{m=1}^{M} \frac{T_m'(f_k)}{\beta T_0(f_k)} \quad (4.12)$$

To simplify the above equation, we introduce the following threshold function [183]:

$$\Delta(x)=\begin{cases}1, x>1\\ 0, \text{otherwise}\end{cases} \tag{4.13}$$

By applying the $\Delta$ function to approximate (4.12), we obtain

$$\frac{TI(f_k)}{\beta T_0(f_k)}=\frac{1}{M}\sum_{m=1}^{M}\Delta\left(\frac{T_m'(f_k)}{\beta T_0(f_k)}\right)=\frac{S_M}{M} \tag{4.14}$$

where $S_M=\sum_{m=1}^{M}\Delta\left(\frac{T_m'(f_k)}{\beta T_0(f_k)}\right)$.

Equation (4.14) can be explained by using the Bernoulli test, in which each scan path is used individually to test feature $f_k$. $S_M$ is the number of trials suggesting that the given object should be attended during visual search. When the scan paths are collected from a group ($M$) of domain experts with similar prior-knowledge and experience, we can assume that there exists an underlying probability $Pr$ of the object being 'hot' or not. Thus, each test can be viewed as a Bernoulli trial with a mean value of $Pr$ and a variance of $Pr(1\text{-}Pr)$. The rate $S_M/M$ has a mean value of $Pr$ and a variance of $Pr(1\text{-}Pr)/M$, and it can be used to estimate the probability:

$$Pr_{estimate}=S_M/M \tag{4.15}$$

By following the De Moivre-Laplace limit theorem, we have

$$Pr\left\{\left|\frac{Pr_{estimate}-Pr}{\sqrt{Pr(1-Pr)/M}}\right|<Z\right\}\approx\int_{-Z}^{Z}\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx \tag{4.16}$$

The confidence interval is therefore,

$$\left[Pr_{estimate}+\frac{Z^2}{2M}\pm Z\sqrt{\frac{Pr_{estimate}-Pr_{estimate}^2}{M}+\frac{Z^2}{4M^2}}\right]\Bigg/\left[1+\frac{Z^2}{M}\right] \tag{4.17}$$

The probability $Pr$ indicates the degree to which the corresponding object in the image is 'hot' or not. It can be used to construct the saliency map for decision support in image understanding. The variance $Pr(1\text{-}Pr)/M$ provides a consistency measure of attention selection and can be used for the pattern comparison of different scan paths.

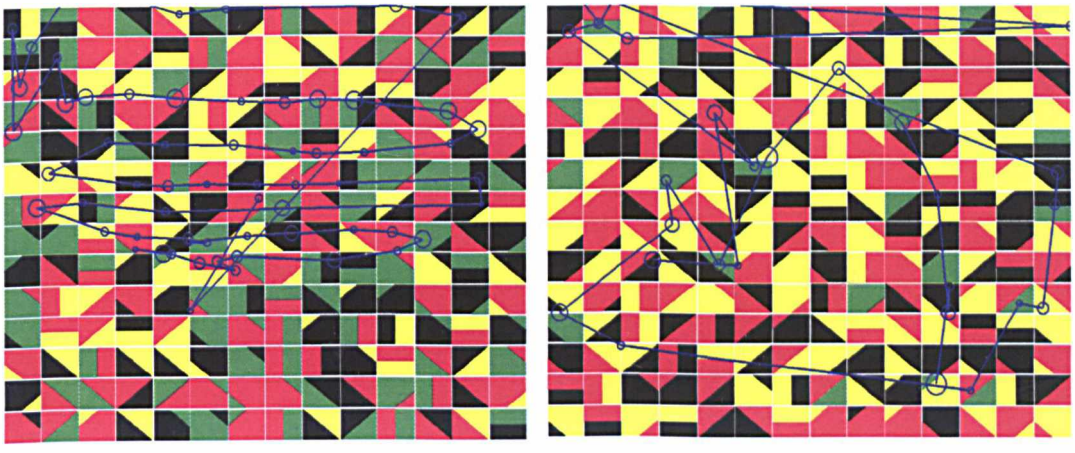
## 4.2 Experiment Setup and Results

### 4.2.1 Experiment Setup

An ASL-504 remote eye tracking system as described in Section 3.1.1 was used to track eye movements in real-time. A chin rest was used to fix the head position during visual search. The images were displayed sequentially on a 24-inch computer monitor (NEC-MultiSync-FP1370; Resolution: 1280×1024 pixels). The effective size of displaying images is 39.4cm in width and 27.6cm in height. The distance from the centre of the screen to the eye is 65.0cm.

### 4.2.2 Two-Colour Conjunction Search

To provide a detailed explanation of the main steps involved in the proposed framework, a two-colour conjunction search experiment was designed. Two example images are shown in Figure 4.2. Each small square in the image contains two different colours. Within each image, there is only one square that contains both green and yellow. The observers were asked to identify the location of the green/yellow block in the shortest time possible. Data recording was stopped as soon as the observer had successfully identified the target.



(a) (b)

**Figure 4.2** A two-colour conjunction search experiment used for illustrating the key steps involved in the proposed framework. Each small square in the image contains two different colours. The observers were asked to identify the square that contains both green and yellow. Data recording was stopped as soon as the observers had successfully identified the target. The size of the circles represents the dwell time of each fixation whereas the lines between fixations indicate the saccadic eye movements. (a) and (b) demonstrate the different search patterns adopted by one observer. The recorded scan paths can be used to extract the underlying strategies used.

As an example, the scan paths for the two test images from one observer are shown in Figure 4.2. To illustrate how the scene-dependent projection bias is eliminated, we assume that the primary features used for visual search in Figures 4.2(a) and (b) are black, red, yellow, and green. As indicated by Figure 4.2(a), this observer adopted a systematic search strategy by sequentially covering the image area with a zig-zag scan path. At an initial glance, the spatial pattern of the scan path of Figure 4.2(b) appeared to be random with no clear indication of the underlying strategy used. Furthermore, none of the examples showed evident spatial fixation clusters. By following (4.6), however, a clear trend started to emerge.

Figures 4.3(a) and (b) illustrate $T(\boldsymbol{f}_k)$, $T_0(\boldsymbol{f}_k)$, and the normalised fixation density function $\Gamma(\boldsymbol{f}_k)$ for the scan paths shown in Figures 4.2(a) and (b), respectively. It is evident from Figure 4.3(a) that the scan path for Figure 4.2(a) gave no preference to any of the colour features. For Figure 4.3(b), however, $\Gamma(\boldsymbol{f}_k)$ clearly indicates that green is where the attention has been focussed upon while searching for the target, despite the random spatial appearance of the scan path. This, in fact, is not surprising. By careful examination of Figure 4.2(b), it can be seen that only a few blocks contain the green colour. Since the target block must have both yellow and green, an efficient strategy would be to visit all green blocks without the need of covering other colours. The scan path shown in Figure

4.2(b) is therefore not randomly formed, but is driven by a clear search strategy. By analysing the transient changes of fixation density, it is also possible to examine the adaptation of search strategy over time. As shown in Figure 4.4(a), the observer searched for yellow at the beginning, and changed to a systematic scanning afterwards, whereas for Figure 4.4(b) the observer seemed to search for yellow and green initially and then adopted the efficient strategy of focusing on green.

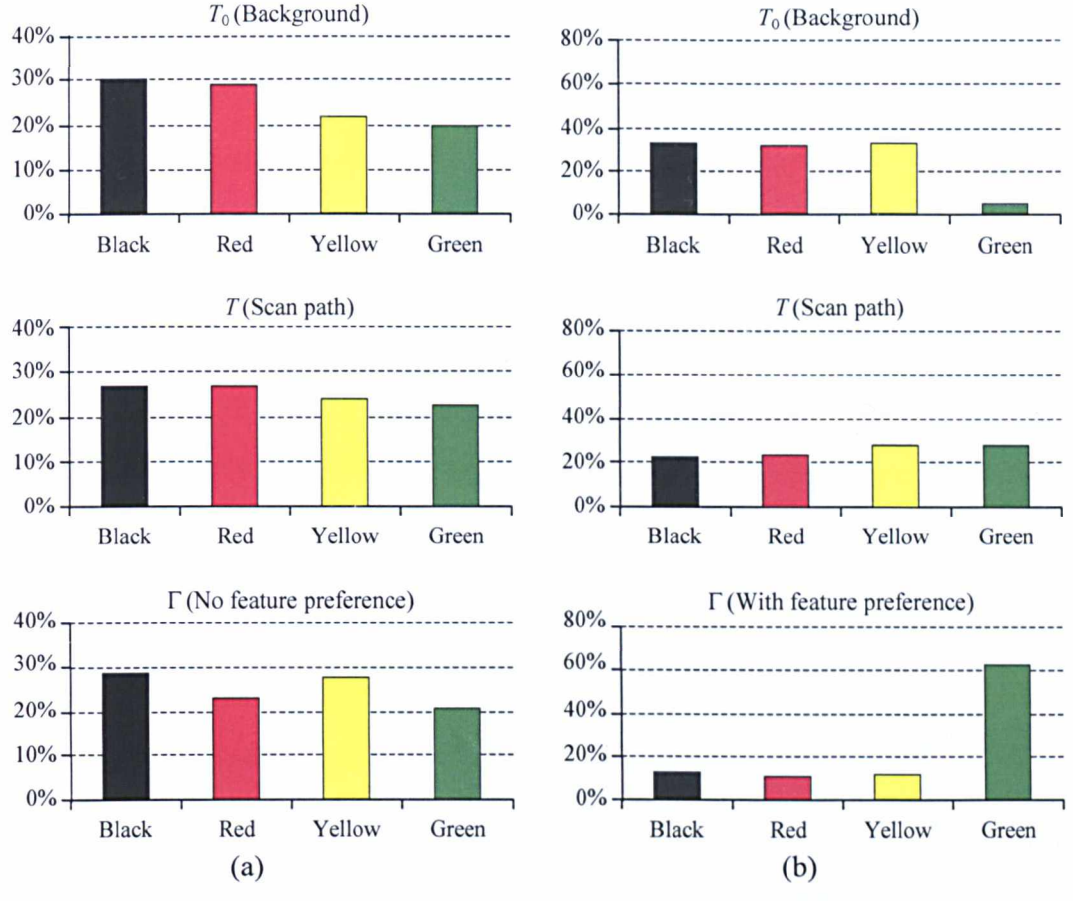



**Figure 4.3** Feature space histograms of $T(f_k)$, $T_0(f_k)$ and $\Gamma(f_k)$ for the scan paths used in Figures 4.2(a) and (b). The fixation density $\Gamma(f_k)$ in (b) has a larger value on green, which indicates that green is where the attention has been focussed upon while searching for the target. These figures also indicate that attention selection in feature space cannot be detected by using total dwell time $T(f_k)$ alone.

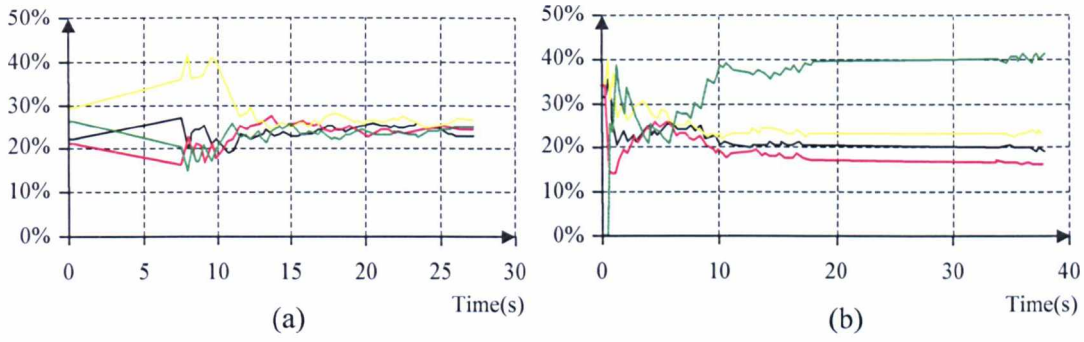



**Figure 4.4** The temporal changes of fixation density function $\Gamma(f_k)$ for the scan paths used in Figures 4.2(a) and (b), illustrating the changes/convergence of visual search strategy.

The same experiments described above were applied to a total of twelve subjects (four females and eight males; mean age=24). Equations (4.15) and (4.17) were used to calculate the estimated probabilities, variance, and 90% confidence intervals. Table 4.1 summarises the results obtained, demonstrating no clear trend of colour preference (all colours<50%) for Figure 4.2(a), and a much evident green colour preference (yellow>50%) for Figure 4.2(b) for this group of subjects studied. The zero probabilities in Table 4.1 suggest that none of the observers gave preference to the corresponding colours.

**Table 4.1** The probabilities of the colours being hot, variance and 90% confidence interval of the two-colour conjunction search experiment for the twelve subjects studied.

| | Figure 4.2(a) | | | | Figure 4.2(b) | | | |
|---|---|---|---|---|---|---|---|---|
| | Black | Red | Yellow | Green | Black | Red | Yellow | Green |
| $Pr_{estimate}$(%) of Equation (4.15) | 2.78 | 0.00 | 36.11 | 19.44 | 0.00 | 2.78 | 0.00 | 83.33 |
| Estimated variance | 0.00075 | 0.00000 | 0.00641 | 0.00435 | 0.00000 | 0.00075 | 0.00000 | 0.00386 |
| 90% confidence interval (%) | 0.62~ 11.58 | 0.00~ 7.03 | 24.31~ 49.86 | 10.88~ 32.30 | 0.00~ 7.03 | 0.62~ 11.58 | 0.00~ 7.03 | 70.83~ 91.15 |

## 4.2.3 Feature Selection and Hot Spot Detection for CT Images of the Lungs

To demonstrate how the proposed technique works in practice, four High Resolution Computed Tomography (HRCT) images of the lungs for a patient with centrilobular emphysema were acquired using a standard high-resolution CT protocol (1.5 mm beam collimation, sharp kernel reconstruction algorithm, on an Imatron C-150-L ultrafast scanner, Imatron, Inc., San Francisco, CA). A DICOM image viewing emulator was used to recreate a normal reporting environment for the observers. The observers can step forward/backward to make comparisons of different slices during the eye tracking

experiment. No image processing was applied to the images shown to the observers. Scan paths derived from two experienced radiologists (A and B) were used to identify common visual texture features used for diagnosis. Figure 4.5 illustrates the HRCT images used in this study. The corresponding scan paths of the two radiologists are superimposed onto the images, demonstrating the idiosyncrasy of their spatial characteristics.

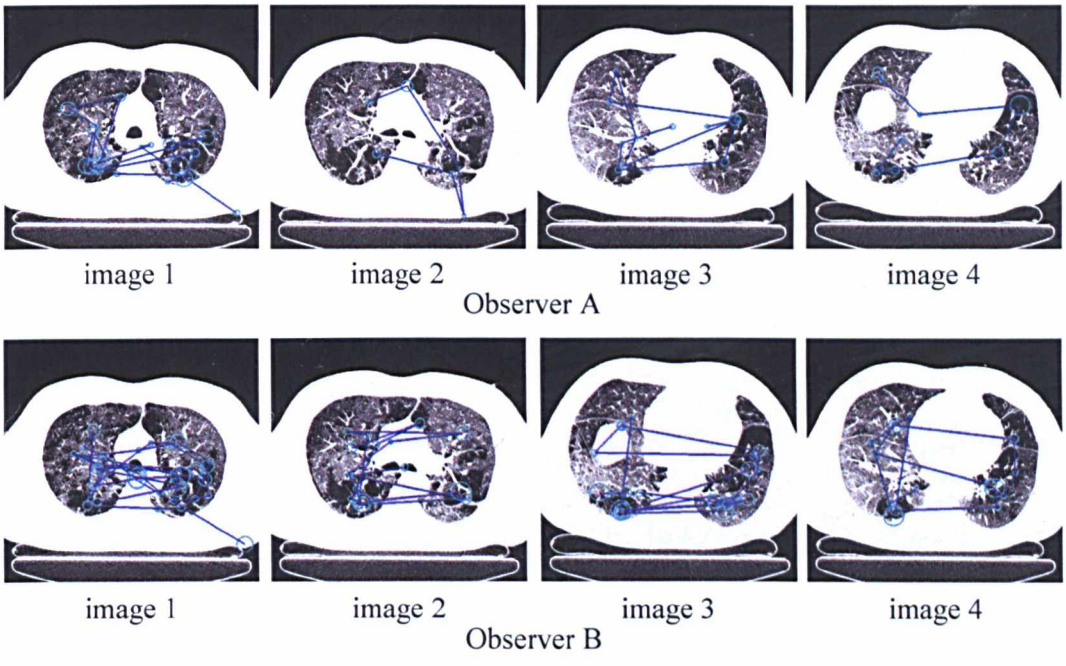



**Figure 4.5** Four HRCT lung images of a patient with centrilobular emphysema and the corresponding visual search scan paths recorded from two experienced radiologists. The size of the circles represents the dwell time of each fixation and the lines between fixations indicate the saccadic eye movements.

The sixteen texture features as listed in Table 3.1 were extracted at every pixel of the images. Each texture feature was quantised into 16 discrete steps. In the analysis of the scan paths, a radius of 1.5° for the foveal field and a probability function modelled with a Gaussian distribution defined by

$$P^{i}_{f_k}(x,y) = \frac{1}{\sqrt{2\pi}\,\delta} exp\left(-\left(a_i(x,y)\right)^2 / 2\delta^2\right) \tag{4.18}$$

were used. In the previous equation, $a_i(x,y)$ represents the visual angle from the centre of the fixation to pixel $<x, y>$, and $\delta$=0.25°. For each feature, the corresponding values for $\Gamma(\cdot)$ and $KL(T, T_0)$ were calculated according to (4.6) and (4.7). The results are shown in Table 4.2. According to the feature selection algorithm described in Section 4.1.3, if we

are to use a single feature to highlight hot spots reflected from a particular scan path, the feature with the highest value of $KL(T, T_0)$ can be selected. Table 4.3 shows the selected features and their corresponding $KL(T, T_0)$ values.

**Table 4.2** *KL* Divergences between $T(\bullet)$ and $T_0(\bullet)$ in the feature space for the scan paths made by two observers when assessing the HRCT images. The highlighted values are the largest values of the columns.

| Observer | A | | | | B | | | |
|---|---|---|---|---|---|---|---|---|
| Scan path | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| mom-mean | 0.80 | 0.80 | 0.62 | 1.10 | 0.22 | 0.30 | 0.53 | **0.43** |
| mom-standard-deviation | 0.49 | 1.23 | 1.03 | 0.45 | 0.34 | 0.67 | 0.94 | 0.26 |
| mom-absolute-deviation | 0.47 | 1.19 | 0.90 | 0.85 | **0.39** | 0.67 | 0.89 | 0.24 |
| mom-skewness | 0.90 | 0.71 | 0.75 | 1.04 | 0.25 | 0.40 | 0.54 | 0.35 |
| mom-kurtosis | 0.07 | 0.47 | 0.30 | 1.15 | 0.21 | 0.42 | 0.31 | 0.09 |
| co-energy | 0.37 | 0.46 | 0.69 | 1.24 | 0.06 | 0.29 | 0.13 | 0.06 |
| co-entropy | 0.17 | 0.20 | 0.34 | **1.47** | 0.11 | 0.75 | 0.17 | 0.12 |
| co-maximum | **1.08** | **1.44** | 0.96 | 1.03 | 0.27 | **0.78** | 0.32 | 0.05 |
| co-contrast | 0.15 | 0.64 | 0.41 | 0.21 | 0.18 | 0.29 | 0.35 | 0.34 |
| co-homogeneity | 0.12 | 0.34 | 0.45 | 1.41 | 0.07 | 0.75 | 0.16 | 0.12 |
| run-spe | 0.92 | 0.74 | **1.17** | 0.65 | 0.09 | 0.43 | **0.96** | 0.29 |
| run-lpe | 0.89 | 1.13 | 1.16 | 0.47 | 0.13 | 0.42 | 0.76 | 0.11 |
| run-glu | 0.85 | 0.84 | 0.75 | 1.06 | 0.07 | 0.45 | 0.15 | 0.09 |
| run-plu | 0.04 | 0.37 | 0.68 | 1.19 | 0.14 | 0.52 | 0.17 | 0.17 |
| entropy | 0.10 | 0.39 | 0.51 | 1.00 | 0.22 | 0.34 | 0.35 | 0.11 |
| fractal | 0.39 | 1.12 | 0.42 | 0.95 | 0.14 | 0.05 | 0.23 | 0.17 |

**Table 4.3** The selected texture features for scan path analysis in 1D feature space and their corresponding *KL* Divergence. Both scan paths 1 and 4 of observer B have rather small *KL* values (highlighted), which indicates that additional features should be considered in determining hot spots during visual search.

| Observer | Scan path | Selected feature | *KL* value |
|---|---|---|---|
| A | 1 | co-maximum | 1.08 |
| | 2 | co-maximum | 1.44 |
| | 3 | run-spe | 1.17 |
| | 4 | co-entropy | 1.47 |
| B | 1 | mom-absolute-deviation | **0.39** |
| | 2 | co-maximum | 0.78 |
| | 3 | run-spe | 0.96 |
| | 4 | mom-mean | **0.43** |

Figure 4.6 demonstrates the back-projection of feature space to spatial regions of interest. The chart illustrates the feature dwell time and fixation density functions for scan path 1 of radiologist A. An optimal threshold $\beta$=1.0 was applied, leading to co-maximum values larger than 6.0 being selected. The highlighted areas on the HRCT image represent salient regions of interest by following the visual search strategy used in Figure 4.5 (observer A, image 1). Similar back-projections can be obtained for other scan paths of this observer.

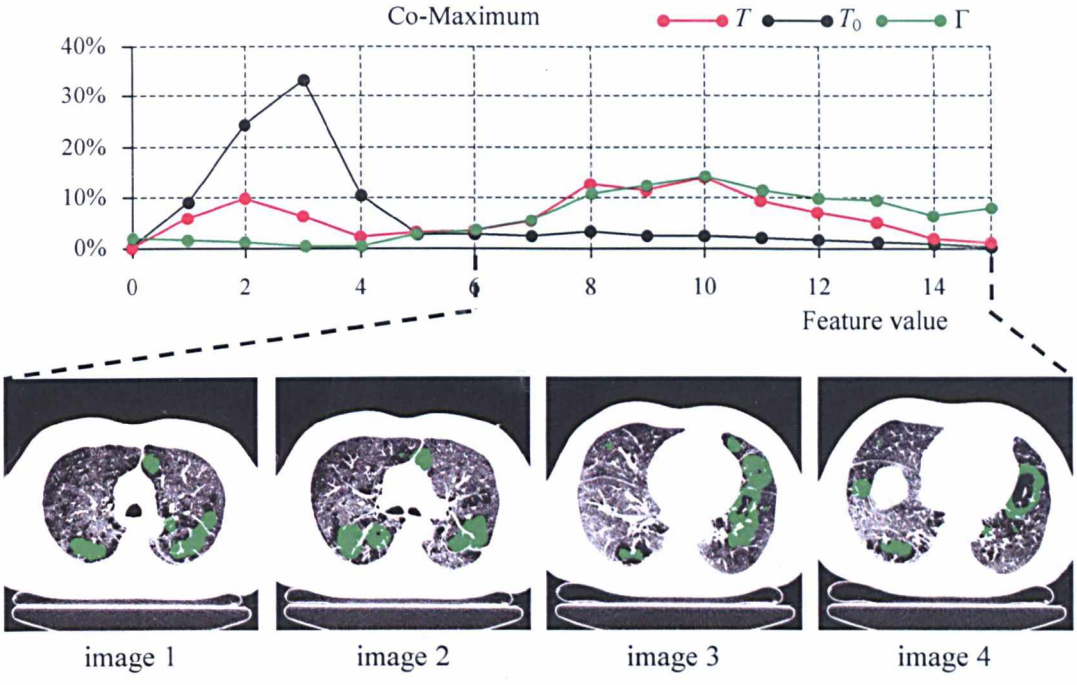



**Figure 4.6** Feature space representation of $T(\cdot)$, $T_0(\cdot)$ and $\Gamma(\cdot)$ corresponding to scan path 1 of radiologist A. The peak of $\Gamma(\cdot)$, with feature values ranging from 6.0 to 15.0, represents hot features and can be back-projected onto the original HRCT lung images, indicating regions of interest where visual attention should be directed towards. The highlighted areas of the four HRCT images are obtained by back-projection of the feature Co-Maximum with value from 6.0 to 15.0.

The use of a single feature for back-projection is simple to implement, however, it may not always be optimal in practice. As shown in Table 4.3, although observer A mainly relied on a single feature for visual search, observer B did not have a clear preference on individual features, especially for images 1 and 4. This indicates that a combination of features was used in the diagnosis. In this case, the forward selection method described in Section 4.1.3 has to be used. The result indicates that the feature pair of absolute deviation and skewness has the largest $KL(T, T_0)$ value both for scan path 1 (0.53) and scan path 4 (0.49). Figure 4.7 demonstrates the 2D fixation density function $\Gamma(\cdot)$ of absolute deviation and skewness for scan path 1 of observer B. After thresholding with $\beta$=1.0, two hot regions can be identified in the feature space. By projecting these two hot regions back to the original HRCT images, the pixels labelled as green and blue indicate the primary regions of interest for this observer. Diagnostically, the green areas represent regions of the lung with decreased attenuation due to centrilobular emphysema and the blue areas are blood vessels and normal lung parenchyma used for cross comparison.

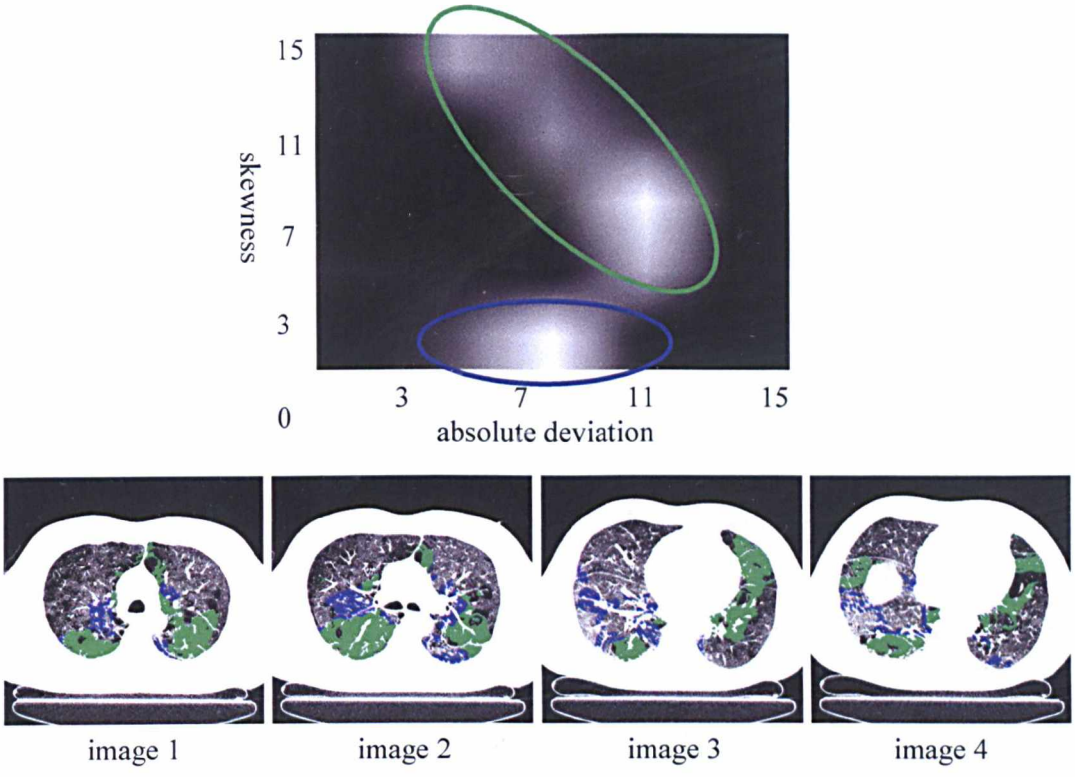



**Figure 4.7** The fixation density function $\Gamma(\cdot)$ of scan path 1 recorded from radiologist B in the feature space formed by texture features of absolute deviation and skewness. The two highlighted regions can subsequently be back-projected to the original HRCT lung images.

Since the hot feature spots derived from each scan path on each image can be used to predict salient regions for other images within the same diagnostic context, a total of eight predicted regions of interest can be derived for each image. Figure 4.8 shows the back-projections for image 1 from the two observers with recorded scan paths for images 1-4, respectively. Equation (4.15) was used to combine the results from the two observers. Figure 4.9 shows the estimated probabilities and variance for the four HRCT images studied. The last row illustrates the derived regions of interest by combining the visual search characteristics of the two observers. It is this type of image that can be used for decision support in medical image understanding.

The selected CT images are different in the shapes of the parenchyma as well as the distribution of the 'black' regions. In a situation like this, traditional spatial domain analysis cannot be directly applied to extract common strategies of visual search even after image registration. The spatial distribution of visual attention can be contrastingly different even when observers use the same strategy to search for diseases. It is evident that the back-projected regions in the above study are highly consistent, which indicates the strength of the proposed technique in robustly extracting common visual behaviours

from different observers, as well as from the same observer, while reading different images within the same context.

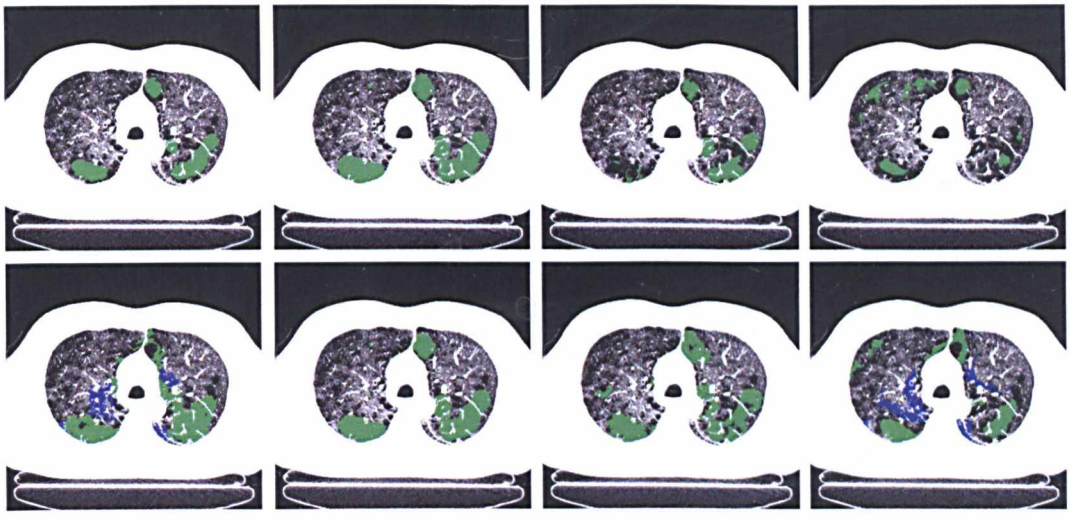
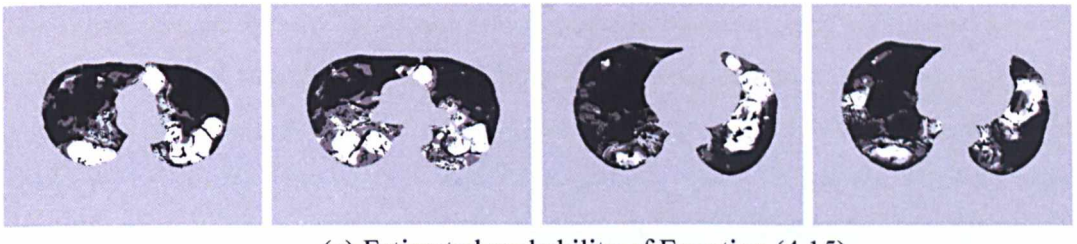
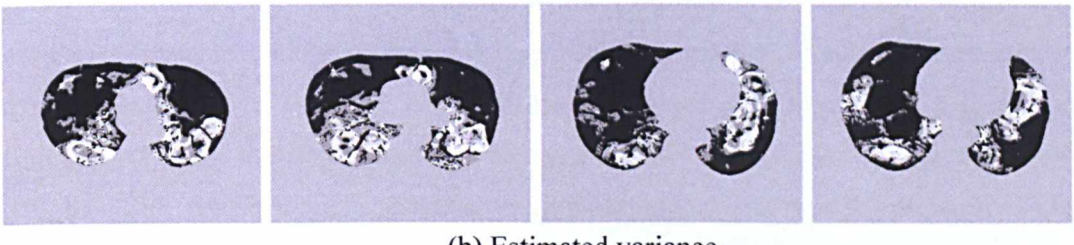


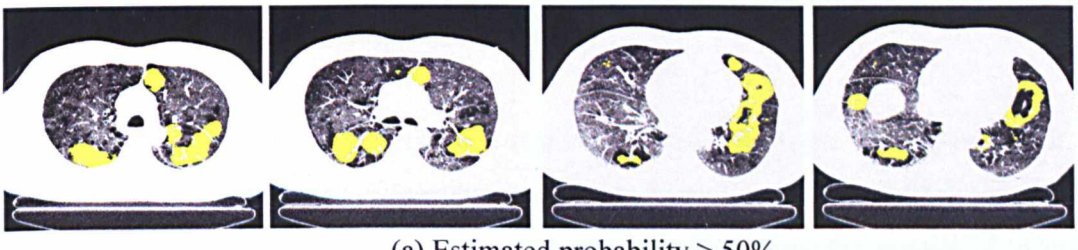
**Figure 4.8** Back-projections for image 1 from the detected hot spots in feature space with scan paths recorded from the two radiologists for images 1-4 respectively.



(a) Estimated probability of Equation (4.15)



(b) Estimated variance



(a) Estimated probability > 50%

image 1 image 2 image 3 image 4

**Figure 4.9** Combined analysis of regions of interest from the two observers by following the Bernoulli trial, where brighter grey shades within the lung parenchyma in (a) and (b) signify higher estimated probabilities and variance, respectively. Those pixels highlighted in the images of (c) have an estimated probability greater than 50%.

## 4.3 Discussion and Summary

With this chapter, we have presented a new framework for identifying salient features from visual search scan paths. It relies on the calculation of feature space fixation density normalised with scene-dependent feature distribution. The method has shown promising strengths in identifying intrinsic visual attention from scan paths that have no evident spatial characteristics. Thus far, the use of sequential coverage and spatial clustering of fixations has been the main method of choice for visual search analysis. In neuroscience, psychology and usability studies, the link between scan path and attention is relatively easy to identify as most experiments are conducted with hypothetical settings. Although the wealth of knowledge accumulated in these fields is invaluable to the current problem of image understanding, there lacks a systematic way of performing feature selection and determining visual attention.

The prevalence of feature led visual search patterns has been long recognised in both normal and skilled visual tasks. In searching for lung nodules in chest radiography, for example, it has been shown that visual attention is usually directed to nodule-specific image properties such as contrast and edge gradient even though the most apparent detection strategy is a systematic scanning such as right-to-left and top-down scan when specific information about the location of nodules is not available [180]. The idiosyncrasy of different observers and the commonly distributed appearance of visual features suggest that the existence of spatially differentiable search patterns is rare. This is further complicated by the fact that in most skilled search tasks, there is generally constant cross-referencing to ancillary and corroborative signs to support the initial observation. The proposed feature space analysis untangles the transitions between fixations and provides an effective means of identifying visual attention.

One of the key elements of the proposed feature space representation of fixation density is the elimination of the scene-dependent projection bias. The accuracy of the proposed model is affected by several factors, such as system calibration, the impact of visual segmentation, the completeness of the feature library and feature distribution estimation and quantisation. In this study, explicit domain knowledge was used to define the texture dimension of the feature space as we removed the chest wall and blood vessels prior to texture extraction. A set of new features must be introduced if the roles of the blood

vessels in visual search are to be modelled. The modelling of the resolution distribution of the fixation field is also crucial. For example, if the size of the image is small compared to the foveal vision field, the derived feature space fixation density function will be significantly attenuated, making it indistinguishable to that of the background bias. In this case, it will be difficult to extract meaningful features that visual search is based upon. Studies on the spatial resolution of visual attention indicate that the resolution of attention is not uniformly distributed over the entire visual field. It is finer in the lower visual field than in the upper field [184]. Furthermore, the effective field of view for each fixation is not constant during visual search; instead, it is affected by the properties of the target and other factors such as the subtlety of the target, item density and homogeneity [185]. Searching for a target in a higher signal-to-noise situation requires a lower resolution and therefore brings about a larger effective size of fixation. As a result, an accurate description of the visual field of fixation should be asymmetric and dynamic.

It should be noted that observers may adopt different visual search strategies in different spatial regions of the image if the feature distributions in these regions are different. It is therefore also necessary to analyse scan paths in spatial domain. One may argue that we can also treat spatial coordinates as a part of the feature vectors. In this case, the proposed method can potentially incorporate both spatial and feature information in a unified framework. The exact details of how this can be done, as well as how to take the order of fixation transitions into consideration, require further investigation. It needs to be pointed out that just using spatial or feature space analysis alone may not be sufficient for a complete understanding of visual attention. The results shown here are a preceding step towards object recognition. The detection of 'hot spots' allows the identification of visual attention, which can be used either for decision support by prompting important areas of interest but without explicitly suggesting the underlying object, or for further steps of object recognition by taking into account other information, such as the order with which all hot spots are covered.

With the steady advancement in digital imaging, there is an increasing demand of developing effective and versatile online decision support systems for improving diagnostic accuracy and overall reproducibility. In the current climate of quality assurance and clinical governance, understanding of the intermediate perceptual steps in the diagnostic process may provide unique information about the way different observers reach their conclusions. In other words, it may be used both as a training tool and potentially as a method of determining whether an individual shows an aberrant or

idiosyncratic approach to interpreting diagnostic images, with the added benefit that corrective training can be gained from the system. Although computerised decision support has been increasingly applied to medical imaging, there has been no generic way of designing such systems. Each application is treated as a new problem, and requires considerable amount of interaction between clinical radiologists and computer scientists in order to identify intrinsic visual features that characterise the medical conditions concerned. The process is further hampered by the fact that visual features are difficult to describe and assimilation of near-subliminal information is cryptic. These drawbacks call for the development of a new framework for a natural and systematic way of gathering knowledge from domain experts. By approaching the problem from observations of how expert visual search takes place, we will also open new possibilities for training and developing self-learning diagnostic decision support systems. The method presented in this chapter represents one important step towards reaching these goals.

# Chapter 5

# Feature Space Analysis of Fixation Sequences

According to Feature Integration Theory [2] introduced in Chapter 2, the features that define a target are integrated for the recognition of a target in a serial attentive stage. Visual search during this stage is affected not only by low-level features but also by the high-level semantic properties of the target [117]. Existing research has shown that the involvement of cognitive strategies is significant for complex tasks [132]. For these tasks, fixation sequences can be crucial in object recognition to establish the relationship between features and resolve ambiguous identities [6, 129]. It has been long recognised that the modelling of fixation sequences is important for understanding the underlying strategies in visual search [7]. In object recognition, for instance, knowledge about the geometric properties of primitives can be essential for the identification of objects. In medical imaging, the cross-checking of ancillary and corroborative features, and the cross-comparison between normal and abnormal regions by radiologists, are common. Despite the differences in scan paths of different observers, the search patterns are rarely random. Observers tend to have repetitive and idiosyncratic ways of inspecting and recognising objects in a given context [129].

The correct identification of transition patterns is important for the understanding of the cognitive process involved in visual search. In an early paper [186], Noton and Stark stated that '*recordings of the points inspected in the scanning of a picture and of the path the eye follows in the inspection provide clues to the process whereby the brain perceives*

*and recognises objects*'. The first-order Markov matrices and string-editing theory have been used to model and compare fixation sequences in the spatial domain [6, 129]. In our early research [5], first-order Markov modelling was also applied to describe the patterns of fixation sequences obtained when radiologists were examining the vascular redistribution of HRCT images. One inherent limitation of the above Markov modelling approaches is the subjective nature involved in the definition of the transition states. To overcome this problem, spatial clustering methods were employed to detect fixation clusters [5, 178]. In practice, however, it is not always possible to form meaningful spatial fixation clusters [187], since visual search can be performed in feature space rather than in spatial domain, in many practical applications.

In this chapter, we propose a method for identifying the patterns of fixation sequences *via* feature-space analysis. With this approach, the meaningful states of fixation sequences are derived from the results of hot spot detection [187], as described in Chapter 4. We also illustrate how to eliminate the projection bias caused by random scanning as well as how to integrate transition patterns obtained from different observers. Both artificial and medical images are used to validate the method and demonstrate its potential value for knowledge gathering in medical imaging.

# 5.1 Material and Methods

## 5.1.1 Feature Space Representation of Fixation Sequences

Features can be discrete and continuous. For continuous features, the technique of hot spot detection [187] can be applied to segment feature space into several regions, and each region can be labelled by a discrete value. Thus, only discrete features are considered in this study. By following the notation used in Chapter 4, we use $F=\{F_1, F_2,\ldots,F_N\}$ to represent the feature value set, where $F_1, F_2,\ldots,F_N$ are all possible feature values. For instance, the colour features used for visual search in Figure 4.2 are black, red, yellow and green. Thus, $F$={black, red, yellow, green} for these images. The texture features listed in Table 3.1 to describe the contents of CT images shown in Figure 4.5 are continuous. By using hot spot detection, two hot spot regions, as shown in Figure 4.7, are obtained. Features within these two hot regions can be labelled with two discrete values, and features outside the two hot regions can be labelled with another value. As a result, feature set $F$ has three discrete values. The advantages of the above quantisation

procedure are that it is an automatic process and each discrete label is related to the contents of the image with regard to visual search. Therefore, it is appropriate to use each of these discrete feature labels as a transition state for fixation sequence analysis.

As in Chapter 4, given a scan path $\xi=<x_k, y_k, t_k>$ $(1\le k\le L)$, the projection $<x_k, y_k>\rightarrow f_i$ indicates that the fixation centred at $<x_k, y_k>$ has a feature value of $f_i$. A probability function $P_k^i(x,y)$ is assigned to the projection $<x_k, y_k>\rightarrow f_i$, which denotes the probability of a feature point $f_i$ being targeted at fixation $<x_k, y_k>$. To take into account all pixels falling within the foveal field of $<x_k, y_k>$, the value of $P_k^i(x,y)$ is approximated by a Gaussian distribution centred at $<x_k, y_k>$. To project fixation transitions into feature space, we define the weighted frequency of transitions between features, *i.e.* given a scan path $\xi=<x_k, y_k, t_k>$ $(1\le k\le L)$,

$$F(f_i,f_j)=\sum_{1\le k<L}\left(\sum_{(x,y)\in\Omega(x_k,y_k)}P_k^i(x,y)\sum_{(x,y)\in\Omega(x_{k+1},y_{k+1})}P_{k+1}^j(x,y)\right) \tag{5.1}$$

where $<x_{k+1}, y_{k+1}>$ is the fixation point immediately after $<x_k, y_k>$; $\Omega(x_k,y_k)$ and $\Omega(x_{k+1},y_{k+1})$ are the foveal fields of fixations $<x_k, y_k>$ and $<x_{k+1}, y_{k+1}>$; $F(f_i, f_j)$ is the weighted frequency of the transitions from feature $f_i$ to $f_j$. Given feature set $F=\{F_1, F_2,\ldots,F_N\}$, $F(f_i, f_j)$ defines a $N\times N$ transition frequency matrix.

## 5.1.2 The Transition Density Matrix

Given feature set $F=\{F_1, F_2,\ldots,F_N\}$, we define a transition density matrix $D$ as:

$$D=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ & \cdots & & \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{bmatrix} \tag{5.2}$$

Each element $d_{ij}$ of $D$ is proportional to the weighted frequency of fixation transitions defined in (5.1). Given a scan path $\xi=<x_k, y_k, t_k>$ $(1\le k\le L)$ and two features $f_i\in F$ and $f_j\in F$,

$$d_{ij}=a_{ij}F(f_i,f_j),\ 1\le i,j\le N \tag{5.3}$$

where

$$d_{ij} \geq 0 \text{ and } \sum_{j=1}^{N} d_{ij} = 1 \tag{5.4}$$

In (5.3), $\alpha_{ij}$ is a normalisation factor depending on the context of the scene. As illustrated in [187], the projection bias is independent of visual search strategies and is only influenced by the relative abundance of different features within the image. As in hot spot analysis [187], we assume that there is no specific transition preference between features during random scanning. As a result, it can be assumed that the transition pattern is a uniform distribution in this case, *i.e.*

$$d_{ij} = d_i^0,\ 1 \leq i, j \leq N \tag{5.5}$$

or

$$a_{ij} F_0(f_i, f_j) = d_i^0,\ 1 \leq i, j \leq N \tag{5.6}$$

where $d_i^0$ is only related to feature $F_i$ and $F_0(f_i, f_j)$ is the weighted frequency of the transitions from feature $f_i$ to $f_j$ in a random scanning.

Since the probability of the transitions from one object to all other objects in the image is equal during random scanning, we have

$$\begin{aligned} F_0(f_i, f_j) &= \sum_{(x_n, y_n) \in image} \left( \left( \sum_{(x,y) \in \Omega(x_n, y_n)} P_n^i(x,y) \right) \sum_{(x_k, y_k) \in image} \left( \sum_{(x,y) \in \Omega(x_k, y_k)} P_k^j(x,y) \right) \right) \\ &= \left( \sum_{(x_n, y_n) \in image} \left( \sum_{(x,y) \in \Omega(x_n, y_n)} P_n^i(x,y) \right) \right) \left( \sum_{(x_k, y_k) \in image} \left( \sum_{(x,y) \in \Omega(x_k, y_k)} P_k^j(x,y) \right) \right) \\ &= \left( \sum_{(x_n, y_n) \in image} \left( \sum_{(x,y) \in \Omega(x_n, y_n)} P_n^i(x,y) \right) \right) T_0(f_j) \end{aligned} \tag{5.7}$$

where

$$T_0(f_j) = \sum_{(x_k, y_k) \in image} \left( \sum_{(x,y) \in \Omega(x_k, y_k)} P_k^j(x,y) \right) \tag{5.8}$$

By substituting (5.6) and (5.7) into (5.3) and using (5.4) to normalise $d_{ij}$, the scene-specific normalisation factor $\alpha_{ij}$ can be eliminated, leading to

$$d_{ij} = \frac{F(f_i, f_j)/T_0(f_j)}{\sum_{j=1}^{N} F(f_i, f_j)/T_0(f_j)} \tag{5.9}$$

If we compare (5.9) with (4.6), it is not difficult to observe the similarity between the two. $T_0(f)$ is used in (4.6) to normalise the total dwell time of a feature in order to remove the projection bias. In (5.9), however, $T_0(f)$ is used to normalise the transition frequency $F(f_i, f_j)$ to remove the projection bias. Both equations can be explained by using the signal detection theory [130], in which a signal is detected by comparing the noise response with the signal-plus-noise response. The ratio between $F(f_i, f_j)$ and $T_0(f_j)$ is the likelihood ratio of the decision theory. $d_{ij}$ is proportional to the transition frequency and has no specific transition preference during random scanning.

### 5.1.3 Feature Space Analysis of Fixation Sequences

#### *1) Feature Extraction and Hot Spot Analysis*

The framework used in this study for fixation sequence analysis is similar to that illustrated in Figure 4.1. But in this case, we are more concerned with the fixation sequence rather than hot spots. The hot spot detection is used for the quantisation of continuous features to determine transition states. After back-projection of these hot regions into the spatial domain, each object of the image obtains a unique label. In practice, feature quantisation can also be performed by using explicit domain knowledge to mark meaningful feature regions in the image.

#### *2) Pattern Integration*

After features are extracted and quantised, Equation (5.9) is applied to calculate the transition density matrix $D$. A large value of $d_{ij}$ suggests that feature $f_j$ is more likely to be

visited immediately after feature $f_i$. When a set of scan paths collected from a group ($M$) of domain experts with similar prior-knowledge and experience is available, it is possible to extract a combined pattern of fixation sequences.

The overall transitions for a group of observers from feature $f_i$ to feature $f_j$ ($1<i, j<N$) are defined as the sum of the transitions of each observer in a given period of time. During this period, the overall transitions from $f_i$ to $f_j$ in a unit time can be estimated as

$$FS(f_i, f_j) = \sum_{m=1}^{M} P_m(f_i) F_m(f_i, f_j) \tag{5.10}$$

where $P_m(f_i)$ is the probability of feature $f_i$ being visited by observer $m$ and $F_m(f_i, f_j)$ is the transition frequency from $f_i$ to $f_j$. By using frequency to estimate $P_m(f_i)$, the above equation becomes

$$\begin{cases} FS(f_i, f_j) = \sum_{m=1}^{M} F_m(f_i) F_m(f_i, f_j) \\ F_m(f_i) = \dfrac{\sum_{1 \le k \le l_m} \left( \sum_{(x,y) \Omega (x_k, y_k)} P_k^i(x, y) \right)}{\sum_{i=1}^{N} \left( \sum_{1 \le k \le l_m} \left( \sum_{(x,y) \Omega (x_k, y_k)} P_k^i(x, y) \right) \right)} \end{cases} \tag{5.11}$$

where $FS(f_i, f_j)$ is a $N$ by $N$ matrix. By normalising each row of the matrix $FS$, an integrated transition frequency matrix can be derived, *i.e.*

$$\begin{aligned} FI(f_i, f_j) &= \frac{\sum_{m=1}^{M} F_m(f_i) F_m(f_i, f_j)}{\sum_{j=1}^{N} \sum_{m=1}^{M} F_m(f_i) F_m(f_i, f_j)} \\ &= \frac{\sum_{m=1}^{M} F_m(f_i) F_m(f_i, f_j)}{\sum_{m=1}^{M} \sum_{j=1}^{N} F_m(f_i) F_m(f_i, f_j)} \\ &= \frac{\sum_{m=1}^{M} F_m(f_i) F_m(f_i, f_j)}{\sum_{m=1}^{M} \left( F_m(f_i) \sum_{j=1}^{N} F_m(f_i, f_j) \right)} \end{aligned} \tag{5.12}$$

Since $\sum_{j=1}^{N} F_m(f_i, f_j) = 1$, we have

$$FI(f_i, f_j) = \sum_{m=1}^{M} C_{i,m} F_m(f_i, f_j) \tag{5.13}$$

or

$$FI = \sum_{m=1}^{M} C_{i,m} F_m \tag{5.14}$$

where $FI$ is the integrated transition frequency matrix, $F_m$ is the transition frequency matrix of observer $m$, and

$$C_{i,m} = \frac{F_m(f_i)}{\sum_{m=1}^{M} F_m(f_i)} \tag{5.15}$$

Accordingly, the likelihood ratio between $FI(f_i, f_j)$ and $T_0(f_j)$ is

$$\frac{FI(f_i, f_j)}{T_0(f_j)} = \sum_{m=1}^{M} C_{i,m} \frac{F_m(f_i, f_j)}{T_0(f_j)} \tag{5.16}$$

Equations (5.14), (5.15) and (5.16) demonstrate that the integrated transition frequency and the likelihood ratio form a linear relationship.

By using the likelihood ratio to detect frequent transitions from one feature to another, we need to test the value of

$$\frac{FI(f_i, f_j)}{\beta T_0(f_j)} = \sum_{m=1}^{M} C_{i,m} \frac{F_m(f_i, f_j)}{\beta T_0(f_j)} \tag{5.17}$$

where $\beta$ is a predefined threshold value.

Similar to the treatment of (4.12) by introducing the Δ function (4.13), we have

$$\begin{cases} HI(f_i, f_j) = \dfrac{FI(f_i, f_j)}{\beta T_0(f_j)} \approx \sum_{m=1}^{M} C_{i,m} \Delta\left( \dfrac{F_m(f_i, f_j)}{\beta T_0(f_j)} \right) = \sum_{m=1}^{M} C_{i,m} H_m(f_i, f_j) \\ H_m(f_i, f_j) = \Delta\left( \dfrac{F_m(f_i, f_j)}{\beta T_0(f_j)} \right) \end{cases} \tag{5.18}$$

where $HI(f_i, f_j)$ has a value between [0.0, 1.0]. It indicates the preference of the transitions from $f_i$ to $f_j$ for this group of observers.

In general, the use of a set of scan paths to model fixation transitions between features is more reliable than using a single scan path. In addition, integrated scan path modelling can reduce the effect of the bias caused by systematic scanning, which can have no functional relevance to the task [89, 187]. Systematic scanning of a specific image could result in certain transition patterns that cannot be eliminated by using Equation (5.9). However, this bias can be attenuated by using scan paths collected from a set of images on which visual features are randomly distributed.

## 5.2 Experimental Results

### 5.2.1 Two-Colour Conjunction Search

The scan paths acquired from the two-colour conjunction search experiment, as shown in Section 4.2.2, are used to demonstrate the above method. In this experiment, observers were asked to identify the location of the green/yellow block in the shortest time possible. The colour features for visual search are black, red, yellow, and green, *i.e.* $F$={black, red, yellow, green}. Equations (5.1) and (5.9) were applied to calculate the transition matrices and transition density matrices for the scan paths in Figures 4.2(a) and (b). A radius of 1.5° for the foveal field of view was used.

For the scan path in Figure 4.2(a), we have

$$F_a = \begin{bmatrix} & black & red & yellow & green \\ black & 22.94 & \underline{\mathbf{30.73}} & 24.33 & 22.00 \\ red & \underline{\mathbf{30.18}} & 28.61 & 18.24 & 22.98 \\ yellow & 34.49 & \underline{\mathbf{41.63}} & 13.96 & 9.92 \\ green & \underline{\mathbf{29.14}} & 28.30 & 28.02 & 14.54 \end{bmatrix} \quad D_a = \begin{bmatrix} & black & red & yellow & green \\ black & 20.07 & 23.65 & \underline{\mathbf{32.31}} & 23.98 \\ red & \underline{\mathbf{27.02}} & 22.54 & 24.80 & 25.64 \\ yellow & 32.95 & \underline{\mathbf{34.99}} & 20.25 & 11.80 \\ green & 25.41 & 21.71 & \underline{\mathbf{37.08}} & 15.80 \end{bmatrix}$$

For the scan path in Figure 4.2(b), we have

$$F_b = \begin{bmatrix} & black & red & yellow & green \\ black & 29.45 & 14.31 & \underline{\mathbf{31.33}} & 24.90 \\ red & 16.50 & 27.76 & 17.11 & \underline{\mathbf{38.63}} \\ yellow & 9.43 & 20.85 & 22.32 & \underline{\mathbf{47.40}} \\ green & 22.28 & \underline{\mathbf{29.73}} & 29.43 & 18.56 \end{bmatrix} \quad D_b = \begin{bmatrix} & black & red & yellow & green \\ black & 13.73 & 7.48 & 14.37 & \underline{\mathbf{64.42}} \\ red & 5.92 & 11.16 & 6.04 & \underline{\mathbf{76.88}} \\ yellow & 2.97 & 7.36 & 6.91 & \underline{\mathbf{82.77}} \\ green & 11.88 & 17.77 & 15.44 & \underline{\mathbf{54.91}} \end{bmatrix}$$

where the highlighted numbers are the largest numbers of each row. It is apparent that the values in the column green of $D_b$ are larger than those in the other columns, indicating the preference of transitions from the current colour to the green colour while searching for the target in Figure 4.2(b). However, none of the values in $D_a$ is outstandingly large, indicating no transition preference for a systematic scanning. These results coincide with the results obtained in Section 4.2.2. The scan path shown in Figure 4.2(b) is not randomly formed but is driven by a clear search strategy. Since a few blocks in Figure 4.2(b) contain the green colour and the target block must have both yellow and green, an efficient strategy would be to visit all green blocks without the need to cover other colours. It should be noted that the search strategy cannot be observed from the transition frequency $F_a$ and $F_b$.

$$HI_a = \begin{bmatrix} & black & red & yellow & green \\ black & 0.57 & 0.06 & \underline{\mathbf{0.66}} & 0.37 \\ red & \underline{\mathbf{0.68}} & 0.09 & 0.66 & 0.43 \\ yellow & 0.22 & 0.34 & \underline{\mathbf{0.83}} & 0.49 \\ green & \underline{\mathbf{0.74}} & 0.18 & 0.57 & 0.41 \end{bmatrix} \quad HI_b = \begin{bmatrix} & black & red & yellow & green \\ black & 0.40 & 0.17 & 0.38 & \underline{\mathbf{0.92}} \\ red & 0.07 & 0.19 & 0.72 & \underline{\mathbf{0.78}} \\ yellow & 0.16 & 0.33 & 0.46 & \underline{\mathbf{1.00}} \\ green & 0.33 & 0.38 & 0.46 & \underline{\mathbf{0.85}} \end{bmatrix}$$

Scan paths derived from a total of 12 subjects (four females and eight males; mean age=24) were used to identify a common transition pattern. Equation (5.18) was applied to calculate *HI* matrices for scan paths in Figures 4.2(a) and (b). Since the prior knowledge about transitions is not available, we set $\beta$=1 empirically in the calculation. As

before, the highlighted values in the matrices are the largest values of each row. Figure 5.1 is the graphical presentation of the matrices.

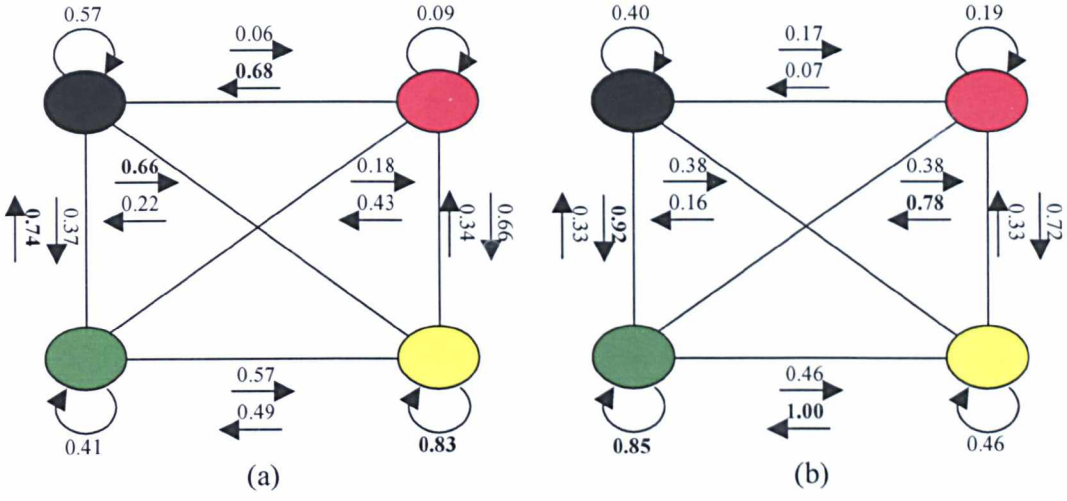



**Figure 5.1** The integrated transition patterns of the scan paths in Figures 4.2(a) and (b). The highlighted values are the largest values outwards each state. The fact that all the values towards green are highlighted in (b) indicates that visual attention is guided to the green colour while searching for the target in Figure 4.2(b).

## 5.2.2 Conjunction Search

### *1) Experiment Design*

Unlike the two-colour conjunction search experiment, where the target is self-defined, real-world search tasks can be more complicated. In Figure 2.9, for example, observers were asked to search for a pair of objects that are the father and son. Since the information about a single object is not sufficient for defining the targets, comparisons between objects are necessary. This situation is common not only in everyday visual search tasks but also in medical imaging where, after an area of interest or abnormality is fixed upon, there often follows a rapid checking of the validity of the observation to exclude technical causes or artefacts and, more importantly, the search for ancillary and corroborative signs to support the initial observation.

In order to investigate how comparisons between objects can affect scan paths and to examine if the proposed method works in this situation, a simplified conjunction search

experiment was constructed. Figure 5.2 shows the basic design of the experiment where each polygon in the image has three features:

- *Size*: small or large.
- *Colour*: black, red, yellow or green.
- *Shape*: 3, 4, 5 or 6 edges.

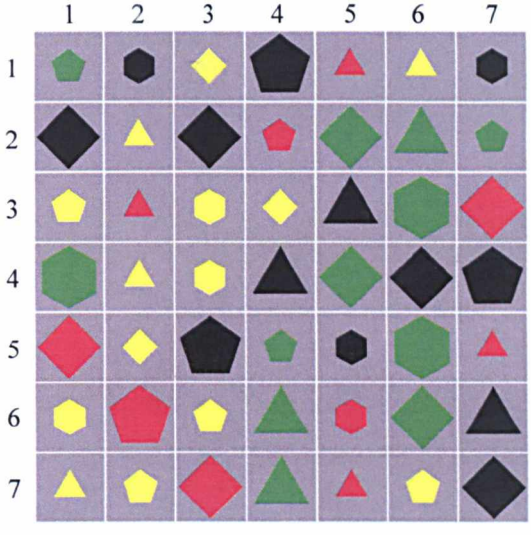
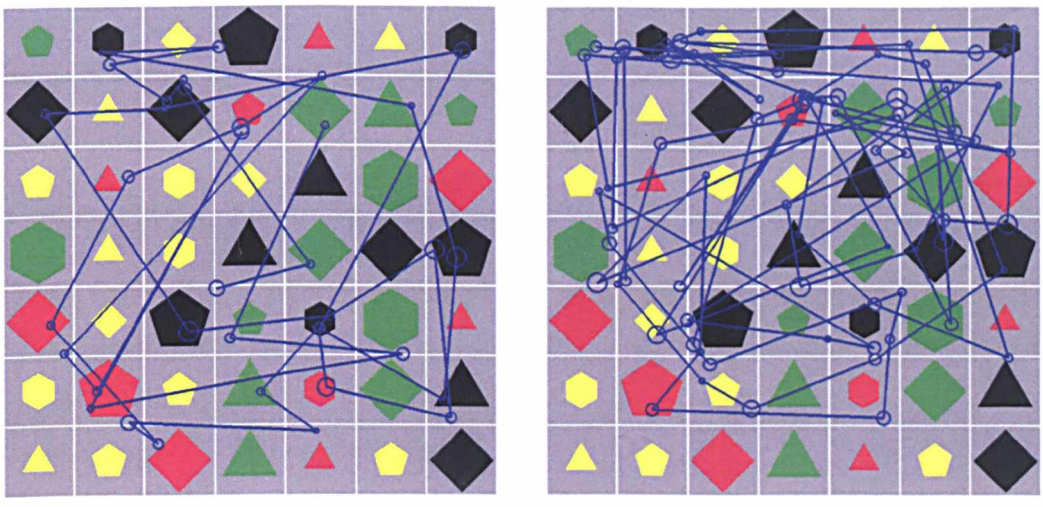



**Figure 5.2** The image designed for examining how comparisons affect scan paths. The polygons can have different sizes (small or large), different colours (black, red, yellow or green) and different shape (3, 4, 5 or 6 edges). The targets are a pair of father and son polygons that have the same colour and shape but are different in size.

The targets are a pair of polygons that have the same colour and shape but are different in size. The coordinates of the targets in Figure 5.2 are (4, 2) and (2, 6). In order to identify common transition patterns, scan paths derived from a total of 12 subjects (four females and eight males; mean age=24) were used in this study.

### *2) Data Analysis*

Figure 5.3 illustrates two scan paths acquired from two observers. At an initial glance, the spatial patterns of the scan paths appear to be chaotic as compared to the scan paths shown in Figure 4.2. None of the examples shows evident spatial fixation clusters.



(a) (b)

**Figure 5.3** Two scan paths acquired from two observers. The size of the circles represents the dwell time of each fixation, and the lines between fixations indicate the saccadic eye movements.

To detect the transition patterns in feature space, the scan paths in different feature spaces (size, colour and shape) were analysed. Equations (5.1) and (5.9) were applied to calculate the transition frequency and transition density matrices. A radius of 1.5° for modelling the foveal field of view was used.

- *Sequence analysis of size*

For the scan path in Figure 5.3(a), we have

$$F_a = \begin{bmatrix} & small & large \\ small & 32.43 & \underline{\boldsymbol{67.57}} \\ large & 45.63 & \underline{\boldsymbol{54.38}} \end{bmatrix} \quad D_a = \begin{bmatrix} & small & large \\ small & 29.95 & \underline{\boldsymbol{70.05}} \\ large & 42.78 & \underline{\boldsymbol{57.22}} \end{bmatrix}$$

For the scan path in Figure 5.3(b), we have

$$F_b = \begin{bmatrix} & small & large \\ small & \underline{\boldsymbol{67.35}} & 32.65 \\ large & \underline{\boldsymbol{53.73}} & 46.27 \end{bmatrix} \quad D_b = \begin{bmatrix} & small & large \\ small & \underline{\boldsymbol{64.75}} & 35.25 \\ large & \underline{\boldsymbol{50.85}} & 49.15 \end{bmatrix}$$

As indicated in the above transition density matrices, the first observer tended to make transitions to large objects, while the second observer tended to make transitions to small objects.

- *Sequence analysis of colour*

For the scan path in Figure 5.3(a), we have

$$F_a = \begin{bmatrix} & black & red & yellow & green \\ black & \underline{\mathbf{74.72}} & 5.99 & 4.34 & 14.94 \\ red & 9.42 & \underline{\mathbf{70.37}} & 5.80 & 14.41 \\ yellow & 41.87 & \underline{\mathbf{46.19}} & 2.80 & 9.15 \\ green & 33.46 & 22.15 & 4.61 & \underline{\mathbf{39.78}} \end{bmatrix} \quad D_a = \begin{bmatrix} & black & red & yellow & green \\ black & \underline{\mathbf{72.96}} & 7.56 & 3.91 & 15.57 \\ red & 7.78 & \underline{\mathbf{75.10}} & 4.41 & 12.71 \\ yellow & 36.76 & \underline{\mathbf{52.41}} & 2.26 & 8.57 \\ green & 30.75 & 26.31 & 3.90 & \underline{\mathbf{39.03}} \end{bmatrix}$$

For the scan path in Figure 5.3(b), we have

$$F_b = \begin{bmatrix} & black & red & yellow & green \\ black & \underline{\mathbf{49.03}} & 11.89 & 20.41 & 18.67 \\ red & 26.98 & \underline{\mathbf{29.40}} & 25.66 & 17.96 \\ yellow & \underline{\mathbf{29.80}} & 24.93 & 16.20 & 29.08 \\ green & 27.19 & 14.01 & 21.80 & \underline{\mathbf{37.00}} \end{bmatrix} \quad D_b = \begin{bmatrix} & black & red & yellow & green \\ black & \underline{\mathbf{47.54}} & 14.90 & 18.23 & 19.33 \\ red & 25.04 & \underline{\mathbf{35.25}} & 21.93 & 17.79 \\ yellow & 27.60 & \underline{\mathbf{29.83}} & 13.82 & 28.75 \\ green & 25.93 & 17.27 & 19.15 & \underline{\mathbf{37.66}} \end{bmatrix}$$

The transition patterns of both observers are similar in terms of colour. They tended to make transitions between the same colours except for yellow. It can be seen that the first observer made more transitions between the same colour, which is probably the reason why the first observer was able to locate the targets faster than the second one, since the targets are defined as the pair of objects with the same colour.

- *Sequence analysis of shape*

For the scan path in Figure 5.3(a), we have

$$F_a = \begin{bmatrix} & 3 & 4 & 5 & 6 \\ 3 & 11.80 & 8.51 & \underline{\mathbf{44.47}} & 35.22 \\ 4 & 11.93 & \underline{\mathbf{51.83}} & 20.47 & 15.78 \\ 5 & 20.79 & \underline{\mathbf{37.07}} & 20.20 & 21.94 \\ 6 & 24.02 & 19.95 & \underline{\mathbf{31.76}} & 24.27 \end{bmatrix} \quad D_a = \begin{bmatrix} & 3 & 4 & 5 & 6 \\ 3 & 9.67 & 7.48 & \underline{\mathbf{42.70}} & 40.15 \\ 4 & 10.51 & \underline{\mathbf{49.03}} & 21.13 & 19.33 \\ 5 & 18.11 & \underline{\mathbf{34.68}} & 20.62 & 26.59 \\ 6 & 20.63 & 18.40 & \underline{\mathbf{31.97}} & 29.00 \end{bmatrix}$$

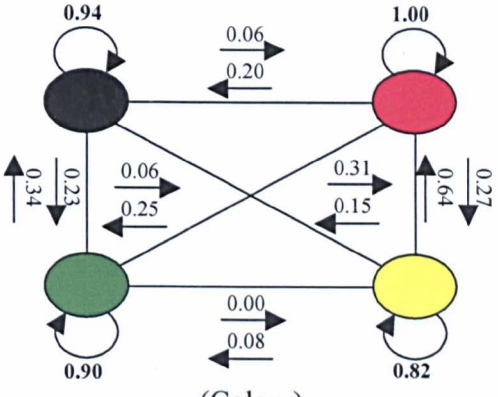
For the scan path in Figure 5.3(b), we have

$$F_b = \begin{bmatrix} & 3 & 4 & 5 & 6 \\ 3 & \underline{\boldsymbol{28.75}} & 24.89 & 21.40 & 24.96 \\ 4 & 12.90 & 20.61 & \underline{\boldsymbol{40.80}} & 25.68 \\ 5 & 19.41 & 20.99 & 25.72 & \underline{\boldsymbol{33.89}} \\ 6 & 11.79 & 17.96 & \underline{\boldsymbol{39.32}} & 30.93 \end{bmatrix} \quad D_b = \begin{bmatrix} & 3 & 4 & 5 & 6 \\ 3 & 24.94 & 23.19 & 21.76 & \underline{\boldsymbol{30.12}} \\ 4 & 10.88 & 18.67 & \underline{\boldsymbol{40.33}} & 30.12 \\ 5 & 16.28 & 18.90 & 25.28 & \underline{\boldsymbol{39.54}} \\ 6 & 9.82 & 16.04 & \underline{\boldsymbol{38.34}} & 35.80 \end{bmatrix}$$

The transition patterns of both observers are very different with regard to shape. No evident coherent pattern can be extracted from the above matrices.

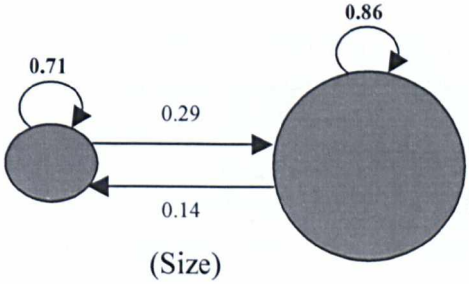
For the 12 subjects studied, Equation (5.18) was used to calculate *HI* matrices. By setting $\beta$=1 empirically in the calculation, the following matrices were derived (the largest value of each row is highlighted):

$$HI_{colour} = \begin{bmatrix} & black & red & yellow & green \\ black & \underline{\boldsymbol{0.94}} & 0.06 & 0.06 & 0.23 \\ red & 0.20 & \underline{\boldsymbol{1.00}} & 0.27 & 0.15 \\ yellow & 0.25 & 0.64 & \underline{\boldsymbol{0.82}} & 0.08 \\ green & 0.34 & 0.31 & 0.00 & \underline{\boldsymbol{0.90}} \end{bmatrix}$$



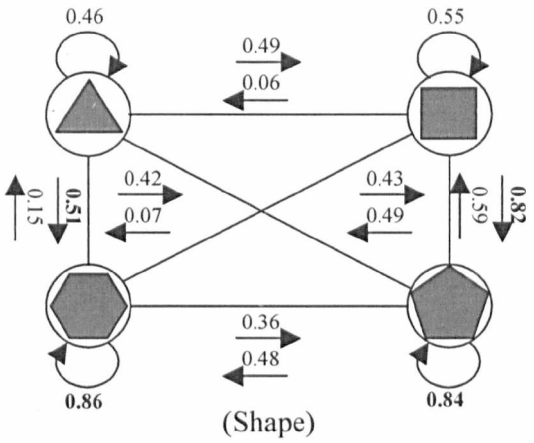

(Colour)

$$HI_{size} = \begin{bmatrix} & small & large \\ small & \underline{\boldsymbol{0.71}} & 0.29 \\ large & 0.14 & \underline{\boldsymbol{0.86}} \end{bmatrix}$$




(Size)

$$HI_{shape} = \begin{bmatrix} & 3 & 4 & 5 & 6 \\ 3 & 0.46 & 0.49 & 0.42 & \underline{\boldsymbol{0.51}} \\ 4 & 0.06 & 0.55 & \underline{\boldsymbol{0.82}} & 0.49 \\ 5 & 0.07 & 0.59 & \underline{\boldsymbol{0.84}} & 0.48 \\ 6 & 0.15 & 0.43 & 0.36 & \underline{\boldsymbol{0.86}} \end{bmatrix}$$



(Shape)

In conclusion, the overall patterns of transitions in this group of observers may be summarised as follows:

- There are significant transitions between objects with the same colour, which indicates that the observers used colour as a primary visual feature. Since colour is a strong 'pop-out' visual feature and the targets are defined as a pair of objects with the same colour, it is reasonable to use colour as a primary feature.
- The observers tended to make transitions between objects with the same size. The logical explanation is that they registered a small (or large) object and went through several other large (or small) objects to locate the targets.
- No coherent pattern can be identified in the sequence of shapes.

## 5.2.3 Fixation Sequence Analysis for CT Images of the Lungs

To apply the above framework to CT image analysis, the CT data as shown in Figure 4.5 were used. For convenience, Figure 5.4 also lists the salient spots extracted by the hot spot detection framework as described in Chapter 4. Diagnostically, the green areas represent regions of the lung with decreased attenuation due to centrilobular emphysema and the blue areas are blood vessels and normal lung parenchyma used for cross comparison. Features within these two hot regions can be labelled with two discrete values $\{blue, green\}$ and features outside the two hot regions can be labelled with another value $\{other\}$. Thus, feature set $F$ has three discrete values and $F=\{blue, green, other\}$.

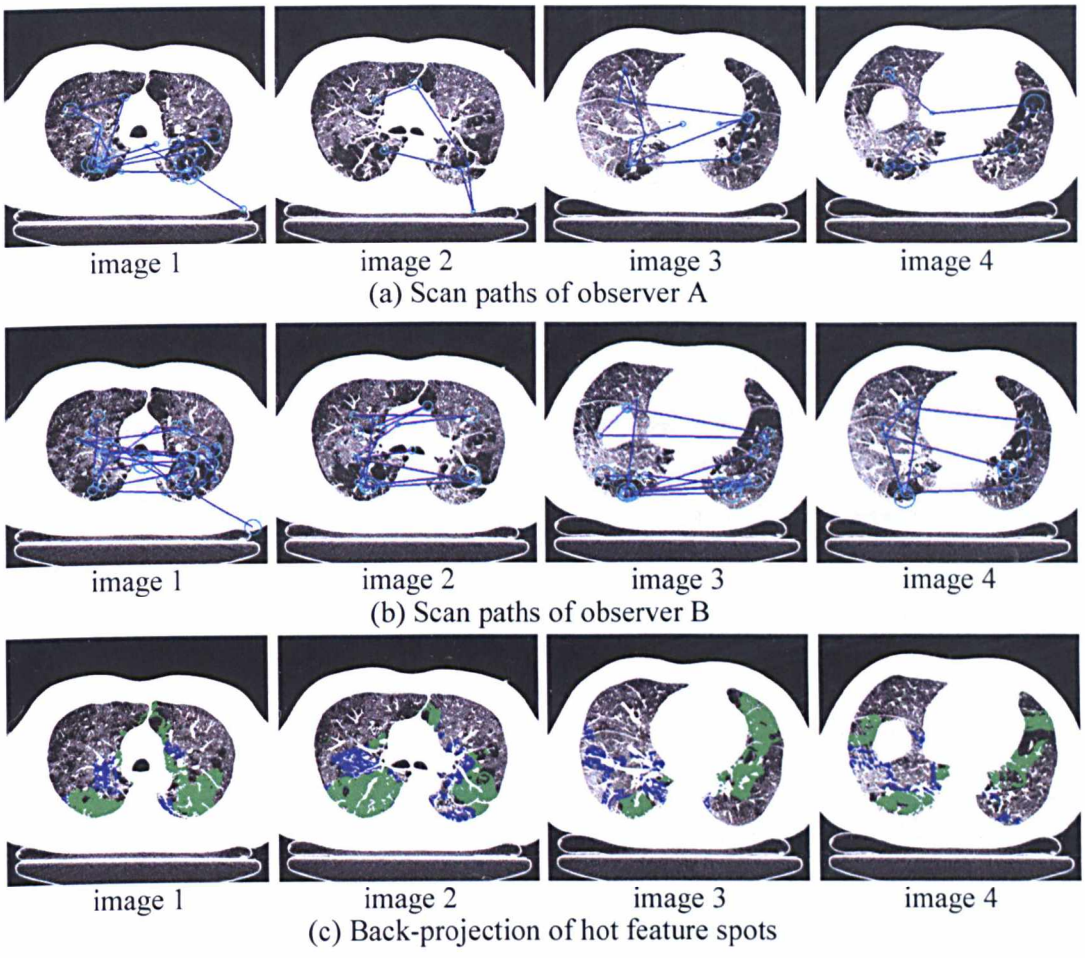
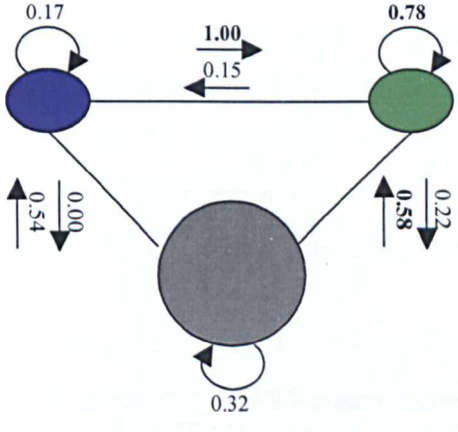



**Figure 5.4** Four HRCT lung images of a patient with centrilobular emphysema and the corresponding visual search scan paths recorded from two experienced radiologists. The two highlighted regions in (c) are obtained by the back-projection of hot feature spots.

For this group of scan paths, Equation (5.18) was used to calculate the *HI* matrix for identifying the common transition pattern. We used the same parameter setting as in the previous experiment. Below are the results for this experiment.

$$HI = \begin{bmatrix} & \textit{blue} & \textit{green} & \textit{other} \\ \textit{blue} & 0.17 & \underline{\boldsymbol{1.00}} & 0.00 \\ \textit{green} & 0.15 & \underline{\boldsymbol{0.78}} & 0.22 \\ \textit{other} & 0.54 & \underline{\boldsymbol{0.58}} & 0.32 \end{bmatrix}$$



The overall patterns of transitions in this group of observers may be summarised as follows:

- There is no strong preference of transitions from 'other' to any feature.
- There are strong preferences of transitions from 'black lung' and 'blood vessels and normal lung parenchyma' regions to 'black lung' regions.

It can be deduced from this pattern that the observers tended to make transitions to 'black lung' areas. It is interesting to see that there are fewer transitions from 'blood vessels and normal lung parenchyma' regions to 'blood vessels and normal lung parenchyma' regions but there are a large number of transitions from 'blood vessels and normal lung parenchyma' to 'black lung', which indicates that visual attention goes back from the blood vessel and normal areas to 'black lung' areas immediately after a cross-comparison has been made.

## 5.3 Discussion and Summary

To achieve optimal visual performance, our visual system tends to select the most informative items within a scene to guide the visual search strategy. How different items are visited can be important in complex search tasks. As a result, the understanding of the strategies about visual attention deployment involves both the detection of regions of interest and the modelling of fixation sequences. In the framework of hot spot detection [187], we have demonstrated how to detect regions of interest in the feature space. In this chapter, we have concentrated on the modelling of fixation transitions between these hot feature regions. We begin by determining meaningful transition states in the feature space, which relies on the identification of features by using the framework of hot spot detection. Spatial fixation sequences are then projected into the feature space to calculate transition matrices. They are subsequently compared with transition patterns obtained from random scanning in order to eliminate the scene-dependent projection bias. We also demonstrate how to integrate transition patterns acquired from a group of scan paths. The strength of the proposed method has been validated by both artificial and medical images.

It should be noted that the accuracy of the proposed model could be affected by the calibration of eye-tracking equipment and the correct computation of scan path fixations. We have illustrated in Section 3.1.2 that recorded gaze coordinates are translated into a

series of fixations by using different criteria, which are 0.5, 1.0 and 1.5 degrees respectively. In this study, these criteria are determined subjectively, and the resultant fixation sequences change with their respective settings. The modelling of the visual field of fixations is crucial in the projection of fixation sequences into feature space. Existing studies indicate that the spatial resolution of attention is not uniformly distributed [184], and the effective size of the visual field of fixations changes with the signal-to-noise ratio of the scene [185]. These asymmetric and dynamic characteristics of the visual field of fixation should be taken into account to further improve accuracy. The focus of this study has been on the first-order patterns of fixation transitions, *i.e.* the likelihood of transitions to the next state is only affected by the present state. In practice, however, the structure beyond the first-order matrices is also important for understanding the visual search behaviours. In the conjunction search experiment, for instance, the first-order modelling is not sufficient to describe transitions from a small object with a given colour to a series of large objects with the same colour. As a result, the complete description of fixation sequences entails the high-order modelling of transition patterns. In addition, comparing patterns obtained from different observers is a key issue in practical applications, such as training in medical imaging. In spatial domain analysis, the minimum string-editing cost has been employed to measure the distance between the spatial patterns of fixation sequences [129]. The method for pattern comparison, as well as the problem of how to incorporate spatial analysis into the framework, requires further investigation.

# Chapter 6

# Bayesian Framework for Feature Selection

The aim of feature selection is to reduce the complexity of an induction system by eliminating irrelevant and redundant features. This technique is becoming increasingly important in the field of machine learning for reducing computational cost and storage, and for improving prediction accuracy. Intuitively, a high-dimensional model is more accurate than a low-dimensional one. However, the computational cost of an inference system increases dramatically with its dimensionality, and therefore we have to balance accuracy and overall computational cost. On the other hand, the accuracy of a high-dimensional model may deteriorate if the model is built upon insufficient training data. In this case, the model would not be able to provide a satisfactory description of the information structures. The amount of training data required to understand the intrinsic structure of an unknown system increases exponentially with its dimensionality. An imprecise description could lead to serious over-fitting problems when learning algorithms are confused by structures brought about by irrelevant features. In a computationally tractable system, less informative features, which contribute little to the overall performance, need to be eliminated. Furthermore, the high cost of collecting a vast amount of sampling data requires efficient selection strategies to remove irrelevant and redundant features.

In machine learning, feature selection methods can often be divided into two groups [188]: wrapper and filter approaches based on the relationship between feature selection

and induction algorithms. The wrapper approach uses the estimated accuracy of an induction algorithm to evaluate candidate feature subsets [189]. By contrast, the filters learn from data and operate independently of any specific induction algorithm. The filter approach evaluates the 'goodness' of candidate subsets based on their information content with regard to target concepts. The filters are not tuned to the specific interactions between the induction algorithm and the information structures embedded in the training dataset. Given enough features, filter-based methods attempt to eliminate features in a way that can maintain as much information as possible about the underlying structure of the data. This makes the technique particularly useful in the understanding of natural systems, such as human visual perception, in which only the most informative features are utilised.

In this chapter, we will describe a filter selection algorithm based on Bayesian theory and Receiver Operating Characteristic (ROC) analysis for feature selection. We demonstrate that likelihood probability plays an important role in the elimination of both irrelevant and redundant features in the algorithm. The experimental results indicate that the proposed method is fast and efficient, and can be used to improve the performance of the classification process. We also explain its potential application to the understanding of behaviours of human visual search.

## 6.1 Feature Relevance and Feature Selection

The techniques for dimensionality reduction have received significant attention in the field of supervised machine learning. Generally speaking, there are two groups of methods: feature extraction and feature selection. In feature extraction, the given features are transformed into a lower-dimensional space, without much loss of information. One of the feature extraction techniques is Principal Component Analysis (PCA) [190], which transforms a number of correlated variables into a number of uncorrelated variables (or *principal components*). For feature selection, however, no new feature is created. The dimensionality is reduced by eliminating irrelevant and redundant features. An irrelevant (or redundant) feature provides no (or no new) information about the target concept [191]. In Bayesian inference, the posterior probability is used for a rational observer to make decisions since it summarises the information available. Consequently, the formal definition of relevance can be based on the conditional independence. [188, 192].

***Definition* 6.1:** Given a set of features $f^{(1)}=\{f_i^{(1)}, 1\le i\le N_1\}$, two sets of features $y$ and $f^{(2)}=\{f_i^{(2)}, 1\le i\le N_2\}$ are conditionally independent or irrelevant if, for any assignment of $y$,

$$Pr(y \mid f^{(1)}) = Pr(y \mid f^{(1)}, f^{(2)}), \text{ whenever } Pr(f^{(1)}, f^{(2)}) \neq 0. \tag{6.1}$$

In the following sections, we use notation $I(y, f^{(2)} \mid f^{(1)})$ to denote the conditional independence of $y$ and $f^{(2)}$ given $f^{(1)}$ . $f^{(1)}$, $f^{(2)}$ and $y$ are assumed to be disjoint without losing generality.

Optimum feature subset selection involves two major difficulties: a search strategy to select candidate feature subsets and an evaluation function to assess these candidates. Figure 6.1 shows a typical model for feature selection:

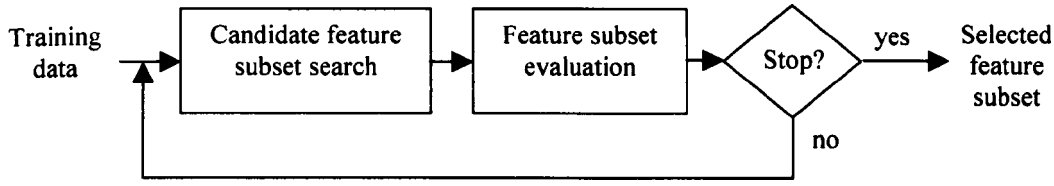



**Figure 6.1** A typical model for feature selection.

The size of the search space for the candidate subset selection is $2^N$, *i.e.* a feature selection method needs to find the best one among $2^N$ candidate subsets given $N$ features. As an example, Figure 6.2 shows the search space for three features:

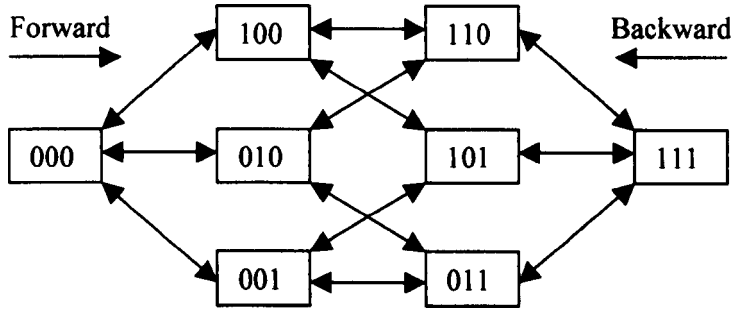



**Figure 6.2** The search space for selecting feature subset from three input features. Each state in the space represents a candidate feature subset. For instance, state 101 indicates that the second feature is not included.

Since the size of the search space grows exponentially with the number of input features, an exhaustive search of the space is impractical. As a result, a heuristic search strategy,

such as the *greedy* search and the *branch and bound* search, becomes necessary [188, 192-194]. *Forward selection* denotes that the search strategy starts with the empty feature set, while *backward elimination* denotes that the search strategy starts with the full feature set. As an example, Koller and Sahami [192] proposed a sequential greedy backward search algorithm to find 'Markov blankets' of features based on the expected cross-entropy evaluation.

The feature selection methods are divided into two groups (*wrapper* and *filter*) according to how the evaluation function is designed. The filters use the information content (typically measured by interclass distance, statistical dependence, and information-theoretic divergences [195]) of the feature subsets for the evaluation. As a result, the filter approaches are independent of any induction algorithm. Kira and Rendell, in their RELIEF [193] algorithm, formulated a weighting method to evaluate the discriminability of each feature by calculating Near-Hit to measure within-concept spread and Near-Miss to measure between-concept separation, whereas the FOCUS [194] algorithm searches for the minimal feature subset *via* a systematic consistency test. But this algorithm is sensitive to noise. On the other hand, the wrappers directly use the predictive accuracy of some induction algorithms to evaluate candidate feature subsets [189]. Statistical techniques such as cross-validation are employed for the purpose of evaluation [196]. Generally speaking, for a given algorithm, wrappers achieve a better accuracy than filters. However, compared to filters, they lack generality and can be computationally demanding. This is because they are directly related to specific induction algorithms, and running these algorithms on a large number of features multiple times can be computationally prohibitive.

Collecting informative training samples from the whole population, and correctly labelling them, are important aspects of feature selection [188]. The cost of labelling can be high, and labelling errors directly affect the performance of feature selection and induction algorithms. In a Bayesian framework, the likelihood probabilities of each class as well as their priors can be estimated independently through controlled experiments. This is useful in practice to avoid direct random sampling from the whole population, which can be a tedious and costly task.

## 6.2 The Theoretical Framework

### 6.2.1 Irrelevance Definition

By using the Bayes rule, for an assignment of $y=a$, Equation (6.1) can be rewritten as

$$\begin{aligned}&\frac{Pr(f^{(1)} \mid y=a)Pr(y=a)}{Pr(f^{(1)} \mid y=a)Pr(y=a)+Pr(f^{(1)} \mid y\neq a)Pr(y\neq a)}\\&=\frac{Pr(f^{(1)},f^{(2)} \mid y=a)Pr(y=a)}{Pr(f^{(1)},f^{(2)} \mid y=a)Pr(y=a)+Pr(f^{(1)},f^{(2)} \mid y\neq a)Pr(y\neq a)}\end{aligned} \tag{6.2}$$

or

$$\left(1+\frac{Pr(f^{(1)} \mid y\neq a)}{Pr(f^{(1)} \mid y=a)}\times\frac{Pr(y\neq a)}{Pr(y=a)}\right)^{-1}=\left(1+\frac{Pr(f^{(1)},f^{(2)} \mid y\neq a)}{Pr(f^{(1)},f^{(2)} \mid y=a)}\times\frac{Pr(y\neq a)}{Pr(y=a)}\right)^{-1} \tag{6.3}$$

Consequently, we can obtain an equivalent definition of relevance.

***Definition* 6.2:** Given a set of features $f^{(1)}=\{f_i^{(1)},1\le i\le N_1\}$, two sets of features $y$ and $f^{(2)}=\{f_i^{(2)},1\le i\le N_2\}$ are conditionally independent or irrelevant if, for any assignment of $y=a$,

$$L(f^{(1)} \,||\, y\neq a, y=a)=L(f^{(1)},f^{(2)} \,||\, y\neq a, y=a), \text{ whenever } Pr(f^{(1)},f^{(2)})\neq 0. \tag{6.4}$$

where $L(f \,||\, y\neq a, y=a)$ is the likelihood ratio,

$$L(f \,||\, y\neq a, y=a)=\frac{Pr(f \mid y\neq a)}{Pr(f \mid y=a)} \tag{6.5}$$

Theoretically, Definitions 6.1 and 6.2 are equivalent. But Definition 6.1 is based on the posterior probability, whereas Definition 6.2 is based on the likelihood ratio. This modification connects feature selection with the performance assessment of decision making. As will be seen in the following sections, a sufficient performance assessment can be achieved by ROC analysis, in which the likelihood ratio plays an essential role.

## 6.2.2 Feature Selection and the Performance of Decision-Making

### 6.2.2.1 Introduction to ROC Analysis

In this section, we examine the relationship between ROC and feature 'irrelevance' to establish the link between feature selection and the performance of decision making. A proper ROC is generated by using the likelihood ratio or its equivalent as the decision variable [197]. Given a pair of likelihoods, the best possible performance of a classifier can be described by the corresponding proper ROC, which can be obtained *via* the Neyman-Pearson ranking procedure by changing the threshold of the likelihood ratio [198]. Given two likelihood distributions $Pr(\boldsymbol{f} \mid y \neq \boldsymbol{a})$ and $Pr(\boldsymbol{f} \mid y = \boldsymbol{a})$ as demonstrated in Figure 6.3(a), the false-alarm and hit rates, according to the Neyman-Pearson procedure, are defined by

$$\begin{cases} P_h = \int_{L(\boldsymbol{f} \mid y \neq \boldsymbol{a}, y = \boldsymbol{a}) > \beta} Pr(\boldsymbol{f} \mid y \neq \boldsymbol{a}) d\boldsymbol{f} \\ P_f = \int_{L(\boldsymbol{f} \mid y \neq \boldsymbol{a}, y = \boldsymbol{a}) > \beta} Pr(\boldsymbol{f} \mid y = \boldsymbol{a}) d\boldsymbol{f} \end{cases} \tag{6.6}$$

where $\beta$ is the threshold, $L(\boldsymbol{f} \| y \neq \boldsymbol{a}, y = \boldsymbol{a})$ is the likelihood ratio as defined by (6.5).

For a given $\beta$, a pair of $P_h$ and $P_f$ can be calculated. When $\beta$ changes from $\infty$ to 0, $P_h$ and $P_f$ change from 0% to 100%. Therefore, the ROC curve is obtained by changing the threshold of the likelihood ratio. Figure 6.3(b) depicts the ROC curve corresponding to the likelihood distributions shown in Figure 6.3(a). In discrete forms, where only some discrete points of $(P_h, P_f)$ are available, straight-line segments are used to connect these points to form a convex hull. Every point on those straight-line segments is realisable by applying a randomised strategy [197, 199]. When misjudgement costs are taken into consideration, minimal maximal-cost points in the ROC space can only be found on the vertices of the convex hull [200].

In a ROC space, the hit rate is the function of the false-alarm rate. The slope at a point on the ROC curve is equal to the associated likelihood ratio. The area under the ROC curve (AUC) is an important measure of discriminability of the two classes described by the two likelihood distributions. The equivalent statistical metric to the AUC is the Wilcoxon

statistics, which was originally designed to estimate the probability of the rank of two random variables [201, 202].

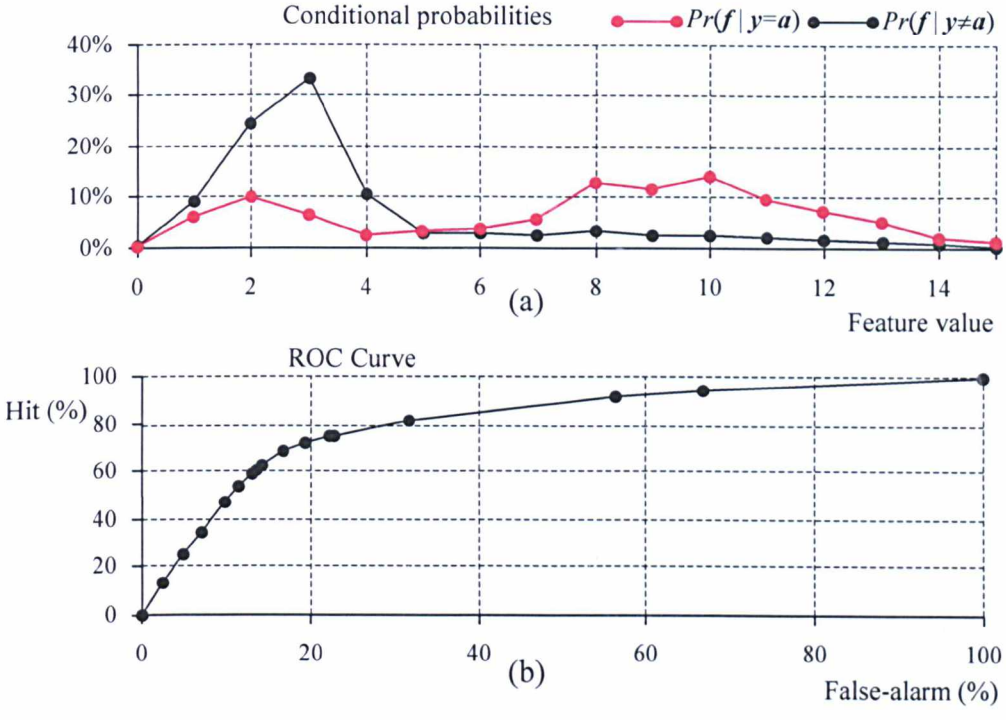



**Figure 6.3** Conditional probabilities and their ROC curve. Two conditional probability functions are plotted in (a). In the ROC space shown in (b), hit rate change s from 0% to 100% when false-alarm rate varies from 0% to 100%.

### 6.2.2.2 ROC and Dimensionality

***Theorem* 6.1:** let $\boldsymbol{f}^{(1)}=\{f_i^{(1)}, 1\le i\le N_1\}$ and $\boldsymbol{f}^{(2)}=\{f_i^{(2)}, 1\le i\le N_2\}$. Given two pairs of likelihood distributions of $Pr(\boldsymbol{f}^{(1)} \mid y\neq a)$, $Pr(\boldsymbol{f}^{(1)} \mid y=a)$ and $Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y\neq a)$, $Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y=a)$, we have two corresponding ROC curves, $ROC(\boldsymbol{f}^{(1)} \parallel y\neq a, y=a)$ and $ROC(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \parallel y\neq a, y=a)$, obtained from the Neyman-Pearson procedure. Then, $ROC(\boldsymbol{f}^{(1)} \parallel y\neq a, y=a)=ROC(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \parallel y\neq a, y=a)$, if and only if

$$L(\boldsymbol{f}^{(1)} \parallel y\neq a, y=a)=L(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \parallel y\neq a, y=a) \tag{6.7}$$

where $L(\boldsymbol{f} \parallel y\neq a, y=a)$ is the likelihood ratio defined in (6.5).

***Proof*:** the proof is given in Appendix A.

***Theorem* 6.2:** let $f^{(1)}$ and $f^{(2)}$ be two feature sets. Given two pairs of likelihood distributions of $Pr(f^{(1)} \mid y\neq a)$, $Pr(f^{(1)} \mid y=a)$ and $Pr(f^{(1)}, f^{(2)} \mid y\neq a)$, $Pr(f^{(1)}, f^{(2)} \mid y=a)$, we have two corresponding ROC curves, $ROC(f^{(1)} \parallel y\neq a, y=a)$ and $ROC(f^{(1)}, f^{(2)} \parallel y\neq a, y=a)$, obtained according to the Neyman-Pearson procedure. Then, $ROC(f^{(1)}, f^{(2)} \parallel y\neq a, y=a)$ is not under $ROC(f^{(1)} \parallel y\neq a, y=a)$ at any point in the ROC space.

***Proof*:** the proof is given in Appendix A.

#### 6.2.2.3 The Definition of Irrelevance Based on Performance and Discriminability

According to Definition 6.2 and Theorem 6.1, we can obtain a new definition of conditional independence and relevance.

***Definition* 6.3:** Given a set of features $f^{(1)}=\{f_i^{(1)}, 1\leq i\leq N_1\}$, two sets of features $y$ and $f^{(2)}=\{f_i^{(2)}, 1\leq i\leq N_2\}$ are conditionally independent or irrelevant if, for any assignment of $y=a$,

$$ROC(f^{(1)}, f^{(2)} \parallel y\neq a, y=a)=ROC(f^{(1)} \parallel y\neq a, y=a) \tag{6.8}$$

where $ROC(f^{(1)}, f^{(2)} \parallel y\neq a, y=a)$ and $ROC(f^{(1)} \parallel y\neq a, y=a)$ are the ROC curves calculated from the Neyman-Pearson procedure given two pairs of likelihood distributions $Pr(f^{(1)}, f^{(2)} \mid y\neq a)$, $Pr(f^{(1)}, f^{(2)} \mid y=a)$ and $Pr(f^{(1)} \mid y\neq a)$, $Pr(f^{(1)} \mid y=a)$, respectively.

Generally speaking, two ROC curves can be unequal when they have the same AUCs. However, Theorem 6.2 holds since $f^{(1)}$ is a subset of $f^{(1)}$ plus $f^{(2)}$. Accordingly, we can obtain another definition of conditional independence and relevance.

***Definition* 6.4:** Given a set of features $f^{(1)}=\{f_i^{(1)}, 1\leq i\leq N_1\}$, two sets of features $y$ and $f^{(2)}=\{f_i^{(2)}, 1\leq i\leq N_2\}$ are conditionally independent or irrelevant if, for any assignment of $y=a$,

$$AUC(f^{(1)}, f^{(2)} \parallel y\neq a, y=a)=AUC\ (f^{(1)} \parallel y\neq a, y=a) \tag{6.9}$$

where $AUC(f^{(1)}, f^{(2)} \parallel y\neq a, y=a)$ and $AUC(f^{(1)} \parallel y\neq a, y=a)$ are the area under the ROC curves calculated from the Neyman-Pearson procedure given two pairs of likelihood

distributions $Pr(\mathbf{f}^{(1)}, \mathbf{f}^{(2)} \mid y \neq a)$, $Pr(\mathbf{f}^{(1)}, \mathbf{f}^{(2)} \mid y = a)$ and $Pr(\mathbf{f}^{(1)} \mid y \neq a)$, $Pr(\mathbf{f}^{(1)} \mid y = a)$ respectively.

Theoretically, Definitions 6.3 and 6.4 are equivalent to Definitions 6.1 and 6.2. But, Definitions 6.3 and 6.4 highlight the effects of feature selection on the performance of decision making and the overall discriminability of a feature set. Definition 6.3 indicates that irrelevant features have no influence on the performance of ideal inference, whereas Definition 6.4 demonstrates that the overall discriminability is not affected by irrelevant features. Starting from Definition 6.3, a forward or backward selection algorithm can be designed to select or eliminate features that cause a larger increase or decrease in accuracy. Similarly, based on Definition 6.4, a selection algorithm can also be implemented *via* discriminability analysis, *i.e.* features that have a large influence on discriminating the target concepts should be selected. In RELIEF [193], for instance, Near-Hit and Near-Miss are calculated to measure within-concept spread and between-concept separation. Definitions 6.3 and 6.4 also indicate that filter approaches select features for ideal inference. This is different from the wrapper approaches since wrapper approaches select features for specific inference algorithms and their ROCs differ from that of the ideal inference due to their own biases.

### 6.2.2.4 Conditional Independence, Discriminability and the ROC

Definition 6.4 reveals the intrinsic relationship between conditional independence and the ROC. It can also be applied to interpret the properties of conditional independence [203]. By using Definition 6.4, we have

$$\mathrm{I}\left(y, \left(\mathbf{f}^{(2)}, \mathbf{f}^{(3)}\right) \mid \mathbf{f}^{(1)}\right) \Rightarrow AUC\left(\mathbf{f}^{(1)}, \mathbf{f}^{(2)}, \mathbf{f}^{(3)} \,||\, y\right) = AUC\left(\mathbf{f}^{(1)} \,||\, y\right) \tag{6.10}$$

According to Theorem 6.2,

$$\begin{cases} AUC\left(\mathbf{f}^{(1)}, \mathbf{f}^{(2)}, \mathbf{f}^{(3)} \,||\, y\right) \geq AUC\left(\mathbf{f}^{(1)}, \mathbf{f}^{(2)} \,||\, y\right) \geq AUC\left(\mathbf{f}^{(1)} \,||\, y\right) \\ AUC\left(\mathbf{f}^{(1)}, \mathbf{f}^{(2)}, \mathbf{f}^{(3)} \,||\, y\right) \geq AUC\left(\mathbf{f}^{(1)}, \mathbf{f}^{(3)} \,||\, y\right) \geq AUC\left(\mathbf{f}^{(1)} \,||\, y\right) \end{cases} \tag{6.11}$$

By combining (6.10) and (6.11), we have

$$\begin{cases} AUC\left(f^{(1)}, f^{(2)}, f^{(3)} \,||\, y\right) = AUC\left(f^{(1)}, f^{(2)} \,||\, y\right) = AUC\left(f^{(1)} \,||\, y\right) \\ AUC\left(f^{(1)}, f^{(2)}, f^{(3)} \,||\, y\right) = AUC\left(f^{(1)}, f^{(3)} \,||\, y\right) = AUC\left(f^{(1)} \,||\, y\right) \end{cases} \tag{6.12}$$

Thus, by using Definition 6.4 again, we have the decomposition property of

$$\begin{aligned} &\mathrm{I}\left(y, \left(f^{(2)}, f^{(3)}\right) \middle| f^{(1)}\right) \Rightarrow \begin{cases} AUC\left(f^{(1)}, f^{(2)} \,||\, y\right) = AUC\left(f^{(1)} \,||\, y\right) \\ AUC\left(f^{(1)}, f^{(3)} \,||\, y\right) = AUC\left(f^{(1)} \,||\, y\right) \end{cases} \\ &\Rightarrow \begin{cases} \mathrm{I}\left(y, f^{(2)} \middle| f^{(1)}\right) \\ \mathrm{I}\left(y, f^{(3)} \middle| f^{(1)}\right) \end{cases} \end{aligned} \tag{6.13}$$

and the weak union property of

$$\begin{aligned} &\mathrm{I}\left(y, \left(f^{(2)}, f^{(3)}\right) \middle| f^{(1)}\right) \Rightarrow \begin{cases} AUC\left(f^{(1)}, f^{(2)}, f^{(3)} \,||\, y\right) = AUC\left(f^{(1)}, f^{(3)} \,||\, y\right) \\ AUC\left(f^{(1)}, f^{(2)}, f^{(3)} \,||\, y\right) = AUC\left(f^{(1)}, f^{(2)} \,||\, y\right) \end{cases} \\ &\Rightarrow \begin{cases} \mathrm{I}\left(y, f^{(2)} \middle| \left(f^{(1)}, f^{(3)}\right)\right) \\ \mathrm{I}\left(y, f^{(3)} \middle| \left(f^{(1)}, f^{(2)}\right)\right) \end{cases} \end{aligned} \tag{6.14}$$

Similarly, for the contraction property, we have

$$\begin{cases} \mathrm{I}\left(y, f^{(3)} \middle| \left(f^{(1)}, f^{(2)}\right)\right) \\ \mathrm{I}\left(y, f^{(2)} \middle| f^{(1)}\right) \end{cases} \Rightarrow \begin{cases} AUC\left(f^{(1)}, f^{(2)}, f^{(3)} \,||\, y\right) = AUC\left(f^{(1)}, f^{(2)} \,||\, y\right) \\ AUC\left(f^{(1)}, f^{(2)} \,||\, y\right) = AUC\left(f^{(1)} \,||\, y\right) \end{cases} \tag{6.15}$$

*i.e.*

$$\begin{aligned} &\begin{cases} \mathrm{I}\left(y, f^{(3)} \middle| \left(f^{(1)}, f^{(2)}\right)\right) \\ \mathrm{I}\left(y, f^{(2)} \middle| f^{(1)}\right) \end{cases} \Rightarrow AUC\left(f^{(1)}, f^{(2)}, f^{(3)} \,||\, y\right) = AUC\left(f^{(1)} \,||\, y\right) \\ &\Rightarrow \mathrm{I}\left(y, \left(f^{(2)}, f^{(3)}\right) \middle| f^{(1)}\right) \end{aligned} \tag{6.16}$$

In conclusion, the conditional independence of features is determined by their intrinsic discriminability, which can be measured by the AUC.

The monotonic property as illustrated by Theorem 6.2 indicates that the overall discriminability of a feature set can be depicted by a graph metaphor. In Figure 6.4, the

overall ability to separate concepts is represented by the union of the discriminability of each feature subset. Each feature subset occupies a portion of the overall discriminability. There can be overlaps between feature subsets. If one feature subset is totally overlapped by other feature subsets, it provides no additional information and can therefore be safely removed without losing the overall discriminability. It needs to be pointed out that the position and area occupied by a feature subset can change when new features are included.

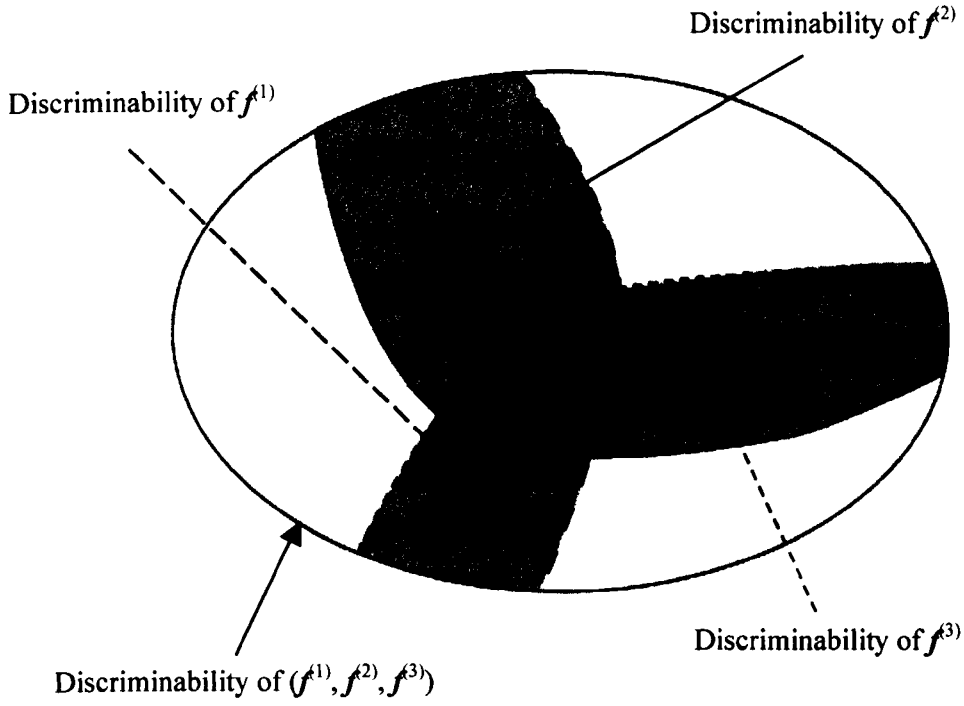



**Figure 6.4** A graphical illustration of the discriminability of features. Each region bordered by an inner curve and the outer circle represents the discriminability of a feature. There can be overlaps between features. The overall discriminability is represented by the area of the region bordered by the outer circle.

### 6.2.2.5 Discriminability Based on the AUC

We denote that the set of possible values of feature $y$ is $\{a_i, i=1, N\}$ and $N$ is the number of classes. $AUC(f \| y \neq a_i, y=a_i)$ denotes the area under the ROC curve of $Pr(f \mid y \neq a_i)$ and $Pr(f \mid y=a_i)$. In this study, the expectation of the AUC is used as an evaluation function:

$$E_{AUC}(f) = E(AUC(f)) = \sum_{i=1}^{N} Pr(y = a_i) AUC(f \| y \neq a_i, y = a_i) \qquad (6.17)$$

In the above equation, the prior probabilities $Pr(y=a_i)$ can either be estimated from data or determined empirically to take misjudgement costs into account. The use of the expected AUC as an evaluation function follows the same principle of sensitivity and specificity. It is not difficult to prove that $E_{AUC}(f^{(1)}, f^{(2)})=E_{AUC}(f^{(1)})$ is equivalent to $AUC(f^{(1)}, f^{(2)} \| y\neq a_i, y=a_i)=AUC(f^{(1)} \| y\neq a_i, y=a_i)$, $\{i=1, N\}$; *i.e.* features in $f^{(2)}$ are irrelevant given features in $f^{(1)}$ according to Definition 6.4. $E_{AUC}(f)$ is also a monotonic function that increases with feature number, and $0.5\leq E_{AUC}(f)\leq 1.0$. For a binary class, $E_{AUC}(f)=AUC(f \| y=a_1, y=a_2)=AUC(f \| y=a_2, y=a_1)$, *i.e.* the calculation of $E_{AUC}(f)$ is not affected by prior probabilities.

To use likelihood distributions to calculate the expected AUC in multiple-class situations, we need to avoid $Pr(f \mid y\neq a_i)$ in (6.17). By using the Bayes rule, we have

$$\begin{aligned} Pr(f \mid y \neq a_i) &= \frac{Pr(y \neq a_i \mid f)Pr(f)}{Pr(y \neq a_i)} \\ &= \frac{\sum_{k=1,N}^{k\neq i} Pr(y=a_k \mid f)Pr(f)}{\sum_{j=1,N}^{j\neq i} Pr(y=a_j)} \\ &= \frac{\sum_{k=1,N}^{k\neq i} Pr(y=a_k)Pr(f \mid y=a_k)}{\sum_{j=1,N}^{j\neq i} Pr(y=a_j)} \\ &= \sum_{k=1,N}^{k\neq i} C_{ki}\, Pr(f \mid y=a_k) \end{aligned} \tag{6.18}$$

where

$$C_{ki} = \frac{Pr(y=a_k)}{\sum_{j=1,N}^{j\neq i} Pr(y=a_j)} (i \neq k) \tag{6.19}$$

By assuming that the decision variable and decision rule for calculating $AUC(f \| y=a_k, y=a_i)$ and $AUC(f \| y\neq a_i, y=a_i)$ are the same [197], we have

$$AUC(f \| y \neq a_i, y=a_i) = \sum_{k=1,N}^{k\neq i} C_{ki} AUC(f \| y=a_k, y=a_i) \tag{6.20}$$

where $AUC(f \| y=a_k, y=a_i)$ represents the area under the ROC curve given two likelihood distributions $Pr(f \mid y=a_k)$ and $Pr(f \mid y=a_i)$ $(i \neq k)$.

Equation (6.20) is used for approximating $AUC(f \| y \neq a_i, y=a_i)$ for multiple-class cases,. By substituting (6.20) into (6.17), we have

$$E_{AUC}(f) = \sum_{i=1}^{N} \left( Pr(y=a_i) \sum_{k=1,N}^{k \neq i} C_{ki} AUC(f \| y=a_k, y=a_i) \right) \tag{6.21}$$

## 6.2.3 Feature Selection Based on ROC Analysis

### 6.2.3.1 Strategies for Feature Selection

By applying the contraction and decomposition properties, we have

$$\begin{cases} I(y, f^{(3)} \mid (f^{(1)}, f^{(2)})) \\ I(y, f^{(2)} \mid f^{(1)}) \end{cases} \Rightarrow I(y, (f^{(2)}, f^{(3)}) \mid f^{(1)}) \Rightarrow \begin{cases} I(y, f^{(3)} \mid f^{(1)}) \\ I(y, f^{(2)} \mid f^{(1)}) \end{cases} \tag{6.22}$$

In the above equation, $I(y, f^{(3)} \mid f^{(1)}, f^{(2)})$ and $I(y, f^{(2)} \mid f^{(1)})$ represent two steps of backward elimination, *i.e.* features in $f^{(3)}$ can be removed when features in $f^{(1)}$ and $f^{(2)}$ are given. This can be immediately followed by another elimination of features in $f^{(2)}$ owing to the existence of features in $f^{(1)}$. $I(y, f^{(3)} \mid f^{(1)})$ indicates that features in $f^{(3)}$ remain irrelevant after features in $f^{(2)}$ are eliminated. As a result, only truly irrelevant features are removed at each iteration by following the backward elimination process. In general, backward elimination is less susceptible to feature interaction than forward selection. This has been found in both wrapper and filter approaches [189, 192]. However, it is impractical for a wrapper approach to apply backward elimination due to the high running cost involved, especially when there are many features. Since removing or adding an irrelevant feature does not change the expected AUC, both backward and forward greedy selection algorithms can be designed to use the expected AUC as an evaluation function.

The proposed backward elimination method is effectively a greedy algorithm. It starts with the full feature set and removes one feature at each iteration. A feature $f_j \in \boldsymbol{f}^{(k)}$ to be removed is determined by using the following equation:

$$f_j = \underset{f_i \in \boldsymbol{f}^{(k)}}{\operatorname{argmin}} \left( E_{AUC}\left(\boldsymbol{f}^{(k)}\right) - E_{AUC}\left(\boldsymbol{f}^{(k)} - \{f_i\}\right) \right) \tag{6.23}$$

where $\boldsymbol{f}^{(k)}=\{f_i,\ 1 \le i \le L\}$ is the temporary feature set after the $k$th iteration.

Estimating the AUC in a high-dimensional space is time-consuming. The accuracy of the estimated likelihood distribution decreases dramatically with the number of features given limited training samples, which in turn introduces ranking error in the AUC estimation [204]. Therefore, an approximation algorithm is necessary to estimate the AUC in a lower-dimensional space.

#### 6.2.3.2 The Approximation Algorithm for Backward and Forward Selection: BFFS algorithm

We explained in Section 6.2.2 that the decrease of the total AUC after removal of a feature $f_i$ is related to the overlap of the discriminability of the feature with other features. In the approximation algorithm, we attempt to construct a feature subset $\boldsymbol{S}^{(k)}$ from the current feature set $\boldsymbol{f}^{(k)}$ and use the degree of discriminability overlap in $\boldsymbol{S}^{(k)}$ to approximate that in $\boldsymbol{f}^{(k)}$. Similar methods have been reported in [192, 205]. A heuristic approach is designed to select $k_s$ features from $\boldsymbol{f}^{(k)}$ that have the largest overlap with feature $f_i$ and we assume that the discriminability overlap of feature $f_i$ with other features in $\boldsymbol{f}^{(k)}$ is dominated by this subset of features. Therefore, the approximation algorithm of backward elimination for selecting $K$ features is as follows:

(a) Let $\boldsymbol{f}^{(k)}$ be the full feature set and $k$ be the size of the full feature set.

(b) Calculate the discriminability differential matrix $M(f_i, f_j); f_i \in \boldsymbol{f}^{(k)}, f_j \in \boldsymbol{f}^{(k)}, f_i \neq f_j$.

$$M(f_i, f_j) = E_{AUC}(\{f_i, f_j\}) - E_{AUC}(\{f_j\}) \tag{6.24}$$

(c) If $k=K$, output $\boldsymbol{f}^{(k)}$.

(d) For $f_i \in \boldsymbol{f}^{(k)}$ $(i=1,k)$

- select $k_s$ features from $\boldsymbol{f}^{(k)}$ to construct a feature subset $\boldsymbol{S}^{(ki)}$. The criterion of the selection is to find feature $f_j$, for which $M(f_i, f_j)$ is the smallest, where $f_j \in \boldsymbol{f}^{(k)}, f_j \neq f_i$.
- calculate $D_{AUC}$.

$$D_{AUC}(f_i) = E_{AUC}(\boldsymbol{S}^{(ki)} + \{f_i\}) - E_{AUC}(\boldsymbol{S}^{(ki)}) \tag{6.25}$$

(e) Select feature $f_d$ with the smallest $D_{AUC}$.

(f) $\boldsymbol{f}^{(k)} = \boldsymbol{f}^{(k)} - \{f_d\}$; $k = k-1$; go to (c).

The approximation algorithm for forward selection is similar:

(a) Let $\boldsymbol{f}^{(k)}$ be empty and $k$ be zero.

(b) Calculate the discriminability differential matrix $M(f_i, f_j)$; $f_i \in \boldsymbol{f}^{(k)}, f_j \in \boldsymbol{f}^{(k)}, f_i \neq f_j$.

$$M(f_i, f_j) = E_{AUC}(\{f_i, f_j\}) - E_{AUC}(\{f_j\})$$

(c) If $k=K$, output $\boldsymbol{f}^{(k)}$.

(d) For $f_i \in \boldsymbol{f}^{(k)}$ $(i=1,k)$

- select $k_s$ features from $\boldsymbol{f}^{(k)}$ to construct a feature subset $\boldsymbol{S}^{(ki)}$. The criterion of the selection is to find feature $f_j$, for which $M(f_i, f_j)$ is the smallest, where $f_j \in \boldsymbol{f}^{(k)}, f_j \neq f_i$.
- calculate $D_{AUC}$.

$$D_{AUC}(f_i) = E_{AUC}(\boldsymbol{S}^{(ki)} + \{f_i\}) - E_{AUC}(\boldsymbol{S}^{(ki)})$$

(e) Select feature $f_d$ with the largest $D_{AUC}$.

(f) $\boldsymbol{f}^{(k)} = \boldsymbol{f}^{(k)} + \{f_d\}$; $k = k+1$; go to (c).

As indicated by the fact that the strong union axiom, $I(y, \boldsymbol{f}^{(2)} \mid \boldsymbol{f}^{(1)}) \Rightarrow I(y, \boldsymbol{f}^{(2)} \mid (\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(3)}))$, does not generally hold for conditional independence [203], irrelevant features can become relevant if more information is given. In other words, feature irrelevance in low-dimensional feature space is different from that in high-dimensional feature space. Theoretically, this fact limits the capacities of all low-dimensional approximation algorithms and forward selection algorithms. In general, the proposed algorithm tends to select features that have large discriminability, and features that provide new information.

## 6.3 Experiment Validation

Both artificial and real-world data, as listed in Table 6.1, were used for algorithm validation. Artificial datasets include Corral [191], LED24 and Monk2-Local [206]. For artificial datasets, the algorithms were trained and validated on well-defined training and test data sets. For real datasets, five-fold cross-validation was used. Table 6.2 illustrates the time cost of backward and forward selection algorithms.

**Table 6.1** Summary of the seven datasets used for validation. Corral, LED24 and Monk2-Local are artificial datasets, and others are real-world datasets. The baseline accuracy for an artificial dataset is the frequency of the majority class on the test dataset, and the baseline accuracy for a real-world dataset is the frequency of the majority class on the whole dataset.

| Dataset | No. of Classes | No. of Features | Training Size | Test Size | Baseline Accuracy |
|---|---|---|---|---|---|
| Corral | 2 | 6 | 32 | 128 | 56.25 |
| LED24 | 10 | 24 | 200 | 3000 | 10.57 |
| Monk2-Local | 2 | 17 | 169 | 432 | 67.13 |
| DNA | 3 | 180 | 3186 | CV | 51.91 |
| Mushroom | 2 | 22 | 8124 | CV | 61.80 |
| Splice | 3 | 60 | 3190 | CV | 51.88 |
| Vote | 2 | 16 | 435 | CV | 61.38 |

**Table 6.2** The CPU time (seconds) for feature selection on a Pentium III (CPU cache: 512KB, CPU frequency: 450MHz, Memory: 384MB) computer under SuSE Linux 7.2 operating system.

| Dataset | No. of Feature Orig./Selected | Forward | | | | Backward | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $k_s$=0 | $k_s$=1 | $k_s$=2 | $k_s$=3 | $k_s$=0 | $k_s$=1 | $k_s$=2 | $k_s$=3 |
| Corral | 6/4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 |
| LED24 | 24/7 | 0.18 | 0.22 | 0.28 | 0.40 | 0.14 | 0.18 | 0.15 | 0.26 |
| Monk2-Local | 17/6 | 0.05 | 0.07 | 0.06 | 0.08 | 0.05 | 0.06 | 0.07 | 0.06 |
| DNA | 180/80 | 276.67 | 288.89 | 311.01 | 344.54 | 275.89 | 277.65 | 279.89 | 290.11 |
| DNA | 180/30 | 277.77 | 286.98 | 301.15 | 327.27 | 276.96 | 278.90 | 281.16 | 292.99 |
| Mushroom | 22/11 | 4.95 | 5.58 | 7.83 | 8.36 | 4.58 | 7.04 | 5.79 | 6.10 |
| Splice | 60/30 | 33.56 | 38.54 | 57.14 | 118.42 | 33.25 | 32.52 | 37.94 | 48.96 |
| Vote | 16/8 | 0.11 | 0.18 | 0.25 | 0.27 | 0.14 | 0.17 | 0.15 | 0.22 |

The motivation of using artificial datasets is to assess the ability of the proposed algorithm in dealing with feature interactions. The three artificial datasets used in this study have strong interactions between relevant features. The concept in the Corral dataset is defined as (A0^A1)V(B0^B1). The other two features in this dataset are 'correlated' and 'irrelevant' features. The correlated feature matches the class label in 75% of instances, making it difficult for the selection algorithms to eliminate. Monk2-Local is a standard training set, containing 11 irrelevant and 6 relevant features to form the following target concept of: exactly two of the relevant features having their first

value. The noise is introduced by reversing the labels of 5% of instances. In the domain of LED24, seven Boolean attributes are used to simulate on-off combinations of LED lights for displaying 10 decimal digits from 0 to 9. Each of the seven Boolean attributes has a 10% chance of reversing its true value. The values of 17 irrelevant features in the LED24 dataset are randomly decided with a 50% chance of being zero or one. Table 6.3 shows the results of the proposed methods.

**Table 6.3** The results of feature selection on artificial datasets. The accuracy is the ratio of the number of relevant features selected against the total number of selected features. The expected AUC in the Table is the expected AUC of selected features.

| Dataset | $k_s$ | Forward | | Backward | |
|---|---|---|---|---|---|
| | | Accuracy | Expected AUC | Accuracy | Expected AUC |
| Corral | 0 | 3/4 | 0.980 | 3/4 | 0.980 |
| | 1 | 3/4 | 0.980 | 3/4 | 0.980 |
| | 2 | 3/4 | 0.980 | 3/4 | 0.980 |
| | 3 | 3/4 | 0.980 | 4/4 | 1.000 |
| | 4 | 3/4 | 0.980 | 4/4 | 1.000 |
| LED24 | 0 | 7/7 | 0.994 | 7/7 | 0.994 |
| | 1 | 7/7 | 0.994 | 7/7 | 0.994 |
| | 2 | 6/7 | 0.994 | 6/7 | 0.994 |
| | 3 | 6/7 | 0.994 | 5/7 | 0.992 |
| | 4 | 5/7 | 0.991 | 5/7 | 0.991 |
| | 5 | 3/7 | 0.989 | 5/7 | 0.993 |
| Monk2-Local | 0 | 2/6 | 0.674 | 2/6 | 0.674 |
| | 1 | 2/6 | 0.680 | 3/6 | 0.816 |
| | 2 | 3/6 | 0.701 | 3/6 | 0.890 |
| | 3 | 2/6 | 0.716 | 5/6 | 0.949 |
| | 4 | 3/6 | 0.705 | 5/6 | 0.968 |
| | 5 | 3/6 | 0.725 | 6/6 | 1.000 |

As indicated in Table 6.3, backward selection is able to detect relevant features by adjusting the value of $k_s$. However, forward selection is not. In the case of the Corral dataset, forward selection always selects the 'correlated' feature because it provides the largest increase in the expected AUC in the first step of selection. This supports the existing observation [189] that backward elimination suffers less from feature interaction than forward selection. It is interesting to see that the expected AUC is consistent with the accuracy, *i.e.* the greater the expected AUC, the higher the accuracy. The strength of using the expected AUC as the evaluation criterion is that it is theoretically justified, as indicated in Section 6.2, and it has clear statistical meaning (equivalent to the Wilcoxon statistics). An expected AUC of 1.0 suggests that class labels can be expressed as a function of selected features, whereas an expected AUC of 0.5 indicates a meaningless selection. Determining a proper value of $k_s$ is also important for both backward and forward selection. In a noisy situation, $k_s$ should not be too large given a limited training dataset. In the case of the LED24 dataset, for instance, larger $k_s$ leads to relatively worse results. In practice, an appropriate $k_s$ could be determined by evaluating different $k_s$ values and their corresponding expected AUCs.

**Table 6.4** A comparison of NB and C4.5 with and without feature selection, where the class probabilities are estimated from data. Accuracy is expressed as the percentage error rate for each classifier.

| Data | Orig./ Selected | $k_s$ | NB | | | C4.5 | | |
|---|---|---|---|---|---|---|---|---|
| | | | Orig. | Fwd | Bckwd | Orig. | Fwd | Bckwd |
| Corral | 6/4 | 0 | 9.38 ± 2.59 | 15.63± 3.22 | 15.63± 3.22 | 18.75± 3.46 | 18.75± 3.46 | 25.00± 3.84 |
| | | 1 | | 15.63± 3.22 | 15.63± 3.22 | | 18.75± 3.46 | 25.00± 3.84 |
| | | 2 | | 15.63± 3.22 | 15.63± 3.22 | | 18.75± 3.46 | 18.75± 3.46 |
| | | 3 | | 15.63± 3.22 | 12.50± 2.94 | | 18.75± 3.46 | 0.00± 0.00 |
| Led24 | 24/7 | 0 | 35.90± 0.88 | 29.47± 0.83 | 29.47± 0.83 | 37.13± 0.88 | 31.20± 0.85 | 31.20± 0.85 |
| | | 1 | | 29.47± 0.83 | 29.47± 0.83 | | 31.20± 0.85 | 31.20± 0.85 |
| | | 2 | | 30.20± 0.84 | 30.37± 0.84 | | 34.33± 0.87 | 35.13± 0.87 |
| | | 3 | | 30.23± 0.84 | 37.03± 0.88 | | 31.90± 0.85 | 36.97± 0.88 |
| Monk2-Local | 17/6 | 0 | 39.35± 2.35 | 39.35± 2.35 | 39.35± 2.35 | 29.63± 2.20 | 32.87± 2.26 | 32.87± 2.26 |
| | | 1 | | 38.43± 2.34 | 38.43± 2.34 | | 32.87± 2.26 | 34.72± 2.29 |
| | | 2 | | 36.57± 2.32 | 36.57± 2.32 | | 32.87± 2.26 | 36.34± 2.32 |
| | | 3 | | 37.04± 2.33 | 36.57± 2.32 | | 32.87± 2.26 | 31.02± 2.23 |
| DNA | 160/80 | 0 | 5.78± 0.28 | 4.93± 0.32 | 4.93± 0.32 | 7.85± 0.75 | 6.78± 0.36 | 6.78± 0.36 |
| | | 1 | | 4.93± 0.18 | 5.05± 0.45 | | 6.75± 0.41 | 6.59± 0.40 |
| | | 2 | | 4.83± 0.30 | 4.39± 0.27 | | 7.03± 0.42 | 6.69± 0.46 |
| | | 3 | | 4.83± 0.22 | 6.81± 0.41 | | 6.72± 0.60 | 6.81± 0.41 |
| DNA | 160/30 | 0 | 5.78± 0.28 | 5.62± 0.27 | 5.62± 0.27 | 7.85± 0.75 | 5.78± 0.42 | 5.78± 0.42 |
| | | 1 | | 5.84± 0.15 | 4.52± 0.22 | | 5.78± 0.32 | 5.52± 0.46 |
| | | 2 | | 5.87± 0.32 | 4.90± 0.42 | | 5.84± 0.46 | 6.31± 0.58 |
| | | 3 | | 6.21± 0.44 | 5.62± 0.45 | | 5.74± 0.59 | 7.03± 0.49 |
| Mushroom | 22/11 | 0 | 0.28± 0.08 | 1.12± 0.18 | 1.12± 0.18 | 0.00± 0.00 | 0.07± 0.05 | 0.07± 0.05 |
| | | 1 | | 0.18± 0.06 | 0.64± 0.05 | | 0.00± 0.00 | 0.00± 0.00 |
| | | 2 | | 0.00± 0.00 | 0.39± 0.06 | | 0.00± 0.00 | 0.00± 0.00 |
| | | 3 | | 2.69± 0.05 | 0.28± 0.03 | | 0.00± 0.00 | 0.07± 0.05 |
| Splice | 60/30 | 0 | 4.64± 0.42 | 3.67± 0.18 | 3.67± 0.18 | 6.08± 0.32 | 5.92± 0.33 | 5.92± 0.33 |
| | | 1 | | 4.04± 0.13 | 3.76± 0.28 | | 5.92± 0.33 | 5.89± 0.33 |
| | | 2 | | 3.70± 0.24 | 3.82± 0.22 | | 5.92± 0.33 | 5.99± 0.34 |
| | | 3 | | 4.58± 0.13 | 3.73± 0.16 | | 6.30± 0.39 | 5.96± 0.31 |
| Vote | 16/8 | 0 | 9.66± 0.86 | 8.97± 1.28 | 8.97± 1.28 | 4.14± 0.46 | 4.37± 0.43 | 4.37± 0.43 |
| | | 1 | | 7.36± 1.18 | 5.52± 0.43 | | 4.37± 0.43 | 4.14± 0.46 |
| | | 2 | | 7.36± 1.18 | 5.52± 0.43 | | 4.37± 0.43 | 4.14± 0.46 |
| | | 3 | | 6.67± 1.05 | 8.51± 1.13 | | 4.14± 0.46 | 4.60± 0.51 |

To demonstrate the performance of the features selected by the proposed algorithm, Naïve Bayes (NB) [207] and C4.5 [208] classifiers were employed. NB assumes that the features are conditionally independent given class labels. It works well when this independence assumption is not severely violated. C4.5 is a tree-style classifier, based on information theory. Information gain ratio is utilised in a top-down process to build a decision tree in C4.5. In order to make the results comparable to other algorithms, we used MLC++2.01 [209] software to perform cross-validation and accuracy assessment. Table 6.4 shows the results of the experiment, in which the prior probabilities of the classes are estimated from the data. After the number of features is reduced, most of the datasets in Table 6.4 have a significant improvement in accuracy. Irrelevant features can provide misleading information when the training data are limited. As a consequence, removing those features can bring about performance improvement. However, for the Corral dataset, the accuracy of NB decreases greatly for both forward and backward selection, although the relevant features can be correctly identified in backward selection

when $k_s$ is larger than 3. The interaction between features in Corral is so strong that the independence assumption is severely violated. Thus, including or eliminating the 'correlated' feature makes a big difference for NB. Its accuracy for Monk2-Local before and after selection is worse than its baseline, showing a high-level interaction between features in this dataset. As demonstrated by other studies [189], the Monk2 dataset is unsuitable for decision trees, such as C4.5. We also observed decreases in the accuracies of C4.5 after feature selection for this dataset.

In Section 6.2.2, we demonstrated that the calculation of the expected AUC is not affected by the prior probabilities of the two classes in a binary class situation. In this section, we have assessed the effect of class probabilities in a multiple-class situation. Instead of estimating the class probabilities from the data, we have used equal class probabilities in the selection process. Table 6.5 illustrates the results.

**Table 6.5** A comparison of NB and C4.5 with and without feature selection, where the class probabilities are set to be equal. Accuracy is expressed as the percentage error rate for each classifier.

| Data | Orig./ Selected | $k_s$ | NB | | | C4.5 | | |
|---|---|---|---|---|---|---|---|---|
| | | | Orig. | Fwd | Bckwd | Orig. | Fwd | Bckwd |
| Corral | 6/4 | 0 | 9.38± 2.59 | 15.63± 3.22 | 15.63± 3.22 | 18.75± 3.46 | 18.75± 3.46 | 25.00± 3.84 |
| | | 1 | | 15.63± 3.22 | 15.63± 3.22 | | 18.75± 3.46 | 25.00± 3.84 |
| | | 2 | | 15.63± 3.22 | 15.63± 3.22 | | 18.75± 3.46 | 18.75± 3.46 |
| | | 3 | | 15.63± 3.22 | 12.50± 2.94 | | 18.75± 3.46 | 0.00± 0.00 |
| Led24 | 24/7 | 0 | 35.90± 0.88 | 29.47± 0.83 | 29.47± 0.83 | 37.13± 0.88 | 31.20± 0.85 | 31.20± 0.85 |
| | | 1 | | 29.47± 0.83 | 30.57± 0.84 | | 31.20± 0.85 | 31.90± 0.85 |
| | | 2 | | 30.20± 0.84 | 30.10± 0.84 | | 34.33± 0.87 | 31.90± 0.85 |
| | | 3 | | 29.90± 0.84 | 36.33± 0.88 | | 31.90± 0.85 | 38.67± 0.89 |
| Monk2-Local | 17/6 | 0 | 39.35± 2.35 | 39.35± 2.35 | 39.35± 2.35 | 29.63± 2.20 | 32.87± 2.26 | 32.87± 2.26 |
| | | 1 | | 38.43± 2.34 | 38.43± 2.34 | | 32.87± 2.26 | 34.72± 2.29 |
| | | 2 | | 36.57± 2.32 | 36.57± 2.32 | | 32.87± 2.26 | 36.34± 2.32 |
| | | 3 | | 37.04± 2.33 | 36.57± 2.32 | | 32.87± 2.26 | 31.02± 2.23 |
| DNA | 160/80 | 0 | 5.78± 0.28 | 4.96± 0.24 | 4.96± 0.24 | 7.85± 0.75 | 6.78± 0.36 | 6.78± 0.36 |
| | | 1 | | 5.15± 0.11 | 5.02± 0.28 | | 6.81± 0.33 | 6.94± 0.38 |
| | | 2 | | 5.08± 0.33 | 4.68± 0.21 | | 7.00± 0.39 | 6.87± 0.44 |
| | | 3 | | 4.74± 0.18 | 4.61± 0.20 | | 6.78± 0.47 | 6.81± 0.39 |
| DNA | 160/30 | 0 | 5.78± 0.28 | 5.62± 0.27 | 5.62± 0.27 | 7.85± 0.75 | 5.78± 0.42 | 5.78± 0.42 |
| | | 1 | | 5.84± 0.15 | 4.52± 0.23 | | 5.78± 0.32 | 5.81± 0.29 |
| | | 2 | | 5.69± 0.15 | 4.68± 0.26 | | 6.03± 0.52 | 6.59± 0.52 |
| | | 3 | | 5.93± 0.25 | 4.58± 0.36 | | 6.40± 0.54 | 6.65± 0.36 |
| Mushroom | 22/11 | 0 | 0.28± 0.08 | 1.12± 0.18 | 1.12± 0.18 | 0.00± 0.00 | 0.07± 0.05 | 0.07± 0.05 |
| | | 1 | | 0.18± 0.06 | 0.64± 0.05 | | 0.00± 0.00 | 0.00± 0.00 |
| | | 2 | | 0.00± 0.00 | 0.39± 0.06 | | 0.00± 0.00 | 0.00± 0.00 |
| | | 3 | | 2.69± 0.05 | 0.28± 0.03 | | 0.00± 0.00 | 0.07± 0.05 |
| Splice | 60/30 | 0 | 4.64± 0.42 | 3.73± 0.24 | 3.73± 0.24 | 6.08± 0.32 | 5.89± 0.33 | 5.89± 0.33 |
| | | 1 | | 3.82± 0.20 | 3.86± 0.30 | | 5.89± 0.33 | 5.92± 0.33 |
| | | 2 | | 3.76± 0.13 | 3.95± 0.24 | | 5.99± 0.31 | 6.02± 0.35 |
| | | 3 | | 4.80± 0.24 | 4.01± 0.25 | | 6.39± 0.38 | 6.21± 0.42 |
| Vote | 16/8 | 0 | 9.66± 0.86 | 8.97± 1.28 | 8.97± 1.28 | 4.14± 0.46 | 4.37± 0.43 | 4.37± 0.43 |
| | | 1 | | 7.36± 1.18 | 5.52± 0.43 | | 4.37± 0.43 | 4.14± 0.46 |
| | | 2 | | 7.36± 1.18 | 5.52± 0.43 | | 4.37± 0.43 | 4.14± 0.46 |
| | | 3 | | 6.67± 1.05 | 8.51± 1.13 | | 4.14± 0.46 | 4.60± 0.51 |

As expected, the binary-class datasets (Corral, Monk2-Local, Mushroom and Vote) have exactly the same results as those shown in Table 6.4, whereas the results of other datasets are slightly different. However, the accuracy of the multiple-class datasets does not deteriorate very much compared with those in Table 6.4. For instance, the Splice dataset has relatively unbalanced class probabilities, which are 24.04%, 24.08% and 51.88% respectively. However, both forward and backward selections achieve significant improvement in accuracy by using equal class probabilities. The results indicate that the likelihood probabilities of features play an essential role in the selection. In practice, it is possible to determine class probabilities empirically based on the prior knowledge of experts, or by assigning non-informative priors.

## 6.4 Conclusions

In this chapter, we presented a filter-based algorithm for feature selection based on Bayesian theory. We have demonstrated the relationship between conditional independence and ideal inference performance, which is described by the ROC curve. We have shown that adding and removing an irrelevant feature will not change the associated AUC. This property, together with the monotonic property of the AUC, defines a theoretical framework that is equivalent to the axiomatic system, consisting of decomposition, weak union and contraction properties [203], for probabilistic dependencies. Based on this framework, the proposed algorithm is designed to provide an accurate, robust and fast feature selection method. The algorithm has shown promising strengths in identifying irrelevant features and improving the accuracy of the classifiers. By using both artificial and real-world datasets for accuracy assessment, we also illustrated the roles of prior and likelihood probabilities in the filter selection.

For a specific induction algorithm, wrapper selection usually achieves a better performance than its filter counterparts in terms of accuracy. However, filter selection is usually faster than wrapper selection, and therefore can serve as a pre-processing step for a wrapper-based method. On the other hand, filter selection provides a more accurate description of the intrinsic structure of an information system, since it is not designed to cater for the bias associated with a specific classifier. As we have demonstrated in Section 6.3, the 'correlated' feature in the Corral dataset will always be selected if a wrapper uses Naïve Bayes as the induction algorithm.

In feature selection, we attempt to reduce the number of selected features and achieve a high expected AUC. This is a dilemma in practice since the AUC is a monotonically increasing function. Given a small training dataset, a high AUC with a large number of selected features could result in over-fitting. Determining the optimal balance is theoretically and practically important. Further investigation is needed to determine the best way to take the factors, such as misclassification cost information and the standard error of the AUC [210], into consideration.

The purpose of this study is to design a theoretical model for understanding human perceptual behaviours [187] such as visual attention, where only the most informative features are utilised. It has been proven in neuroscience, psychology and usability studies that attention selection results from the interaction between top-down and bottom-up visual activities. In the next chapter, we will demonstrate that this interaction can be integrated into the proposed Bayesian framework.

# Chapter 7

# Bayesian Modelling for Visual Feature Selection and Attention Deployment during Visual Search

As described in previous chapters, the need to filter out irrelevant information in human visual search originates from the limited capability of the brain. Early selection models, such as the Early Filtering Model [68] and the Attenuation Model [69], suggest that whole stimuli are analysed by pre-attentive detection, and only a subset of the incoming sensory information is selected [108]. Current physiological studies demonstrate that visual attention affects neural responses to visual stimuli by enhancing the activity in stimulated visual cortical regions and suppressing the unstimulated regions [63, 66]. Therefore, only attended features are enhanced during visual search. Computational modelling of attention deployment usually involves the estimation of visual saliency to suggest promising regions by signifying the likelihood of target presence at each location of the visual scene [1, 2, 12, 67]. In general, computation in a high-dimensional feature space is time-consuming, and the accuracy can decrease dramatically with the number of features [188, 192]. As a consequence, feature selection in visual search is important not only for the better understanding of human visual search but also for the reduction of computational costs in practical applications.

It has been widely accepted that both task-driven top-down activity and task-independent bottom-up information are associated with the deployment of visual attention to direct our

gaze towards objects of interest in visual environments [1]. In saliency-based models, for instance, the saliency map was computed based on the extraction of visual features, such as colour, orientation, motion and texture at each location of the image [2]. In a pure bottom-up model proposed by Itti *et al.* [211], a hierarchical centre-surround mechanism was used to enhance locations with local features different from their neighbourhoods. On the other hand, Moghaddam *et al.* [212] used the information about the target appearance in their statistical framework to calculate the maximum likelihood (ML) in a high-dimensional space. An eigenspace decomposition technique was applied to the efficient estimation of probability density. In Guided Search Model [67], the activation of pre-attentive features was weighted by top-down factors to create an activation map for the prediction of changes in visual attention. Although saliency-based models can explain the spatial deployment of visual attention, they lack a systematic way of selecting relevant visual features.

In this chapter, we propose a computational framework for modelling visual feature selection and attention deployment based on Bayesian theory combined with other biologically-plausible architectures [67, 108, 113]. In the proposed method, visual input is decomposed into visual features in the early stage of visual processing, as suggested by Feature Integration Theory [2]. Visual attention is then directed towards objects in the scene by using feature space representations of both bottom-up (image-based saliency cues) and top-down (task-dependent cues) visual features. We explore the possibility of incorporating human visual search behaviours detected by eye movement analysis for scan path simulation. The focus of this study is on optimal visual feature selection, which we believe is also a consequence of the interaction between top-down and bottom-up processes. Instead of using all visual features to construct a feature space for saliency maps, our system uses a subset of most relevant features. Both artificial and real-world images are used to validate the method and demonstrate its potential for training and knowledge gathering in image understanding.

# 7.1 Bayesian Modelling of Visual Attention

## 7.1.1 Bayesian Modelling

In a statistical framework, the efficient search for a target requires the computation of posterior probability $Pr(O=T \mid \boldsymbol{f}, \boldsymbol{I})$ [212, 213], which is the probability of an object $O$ being target $T$ in image $\boldsymbol{I}$ given a set of visual features $\boldsymbol{f}$. By using Bayes' rule, we have

$$Pr(O=T \mid \boldsymbol{f}, \boldsymbol{I}) = \frac{1}{Pr(\boldsymbol{f} \mid \boldsymbol{I})} Pr(\boldsymbol{f} \mid O=T, \boldsymbol{I}) Pr(O=T \mid \boldsymbol{I}) \tag{7.1}$$

where

- $1/Pr(\boldsymbol{f} \mid \boldsymbol{I})$ (the bottom-up factor): is a scene-dependent feature distribution defining the task-independent saliency of visual features. This formula indicates that infrequent features tend to 'pop-out' of the visual scene, and this has been bolstered by evidence from recent studies on visual attention and applied in the computational modelling of bottom-up visual activities [67, 108, 113]. It has been reported that visual search efficiency is related to stimulus dissimilarity [214], target eccentricity [1] and the ratio between different types of distractors [215]. The formation of the saliency map in the retina, lateral geniculate nucleus and primary visual cortex involves a hierarchical centre-surround mechanism [211] that makes regions that are different from their surround salient.

- $Pr(\boldsymbol{f} \mid O=T, \boldsymbol{I})$ (the top-down factor): is the likelihood of the features in image $\boldsymbol{I}$ when the target is presented. This factor describes the target-dependent information and can also be related to the knowledge and expectation of observers. Top-down control in visual search has been demonstrated in psychophysical experiments in which increased neural activities were recorded in specific cortical regions, such as the frontal, parietal and primary visual cortices [46, 48-50]. In computational models [67], top-down modulation has been applied to enhance the saliency of features belonging to the target object and suppress that of others. This is important for efficient visual search, especially when the target is camouflaged by distractors and bottom-up saliency alone is insufficient to maintain the efficiency. Since the exact appearance of the target in an input image is

usually uncertain to observers before the target is identified, it is possible to approximate this factor as:

$$Pr(f \mid O=T, I) \approx Pr(f \mid O=T) \tag{7.2}$$

This approximation is important for allowing dissociation of the specific image $I$ and target features, indicating that the target appearance is determined by its properties only.

- $Pr(O=T \mid I)$ (the prior probability of the object $O$ being the target in the input image $I$): is a factor employed to model global contextual information about the past experience of searching the target in similar environments [213]. This is based on the evidence that familiar scenes trigger previously successful search strategies in a non-volitional way [216]. It can also be utilised to describe the accumulated knowledge or experience in the current search. For instance, the principle of inhibition of return [1, 47] summarises the fact that the expectation of an object being the target becomes lower after it has been attended. This principle is useful in computational models to prevent visual attention from attending eternally to the location with the maximum saliency [108].

In summary, the activation of an object is determined by both its top-down and its bottom-up factors. By defining

$$\begin{cases} S(f \mid I) = \dfrac{1}{Pr(f \mid I)} \\ A(f \mid O=T, I) = S(f \mid I) Pr(f \mid O=T, I) \approx S(f \mid I) Pr(f \mid O=T) \end{cases} \tag{7.3}$$

Equation (7.1) becomes

$$\begin{aligned} Pr(O=T \mid f, I) &= S(f \mid I) Pr(f \mid O=T, I) Pr(O=T \mid I) \\ &= A(f \mid O=T, I) Pr(O=T \mid I) \end{aligned} \tag{7.4}$$

where $S(f \mid I)$ defines the saliency of an object with features $f$ in image $I$, and $A(f \mid O=T, I)$ defines the activation of an object with features $f$ in image $I$ in searching for target $T$.

Figure 7.1 summarises the above Bayesian modelling for visual attention selection. The input image is decomposed in a parallel fashion into visual features, among which relevant features are selected. The saliency map and top-down control are then computed

in the selected feature space. The saliency map drives visual attention to regions with rare or 'pop-out' features, whereas top-down activation directs visual attention towards regions sharing similar features of the target. The bottom-up saliency is modulated with the top-down bias into a coherent activation map, based on which the input image is sequentially attended to in an order of decreasing activation. The principle of inhibition of return is coded into the selection prior, which is updated constantly during visual search and is determined by the global contextual information of the input image together with the previous searching experience of observers. The model shown in Figure 7.1 is similar to the frameworks previously proposed by Torralba [213] and Koch *et al.* [108, 113] except for feature selection. Instead of using all visual features, our method uses a subset of features that are relevant.

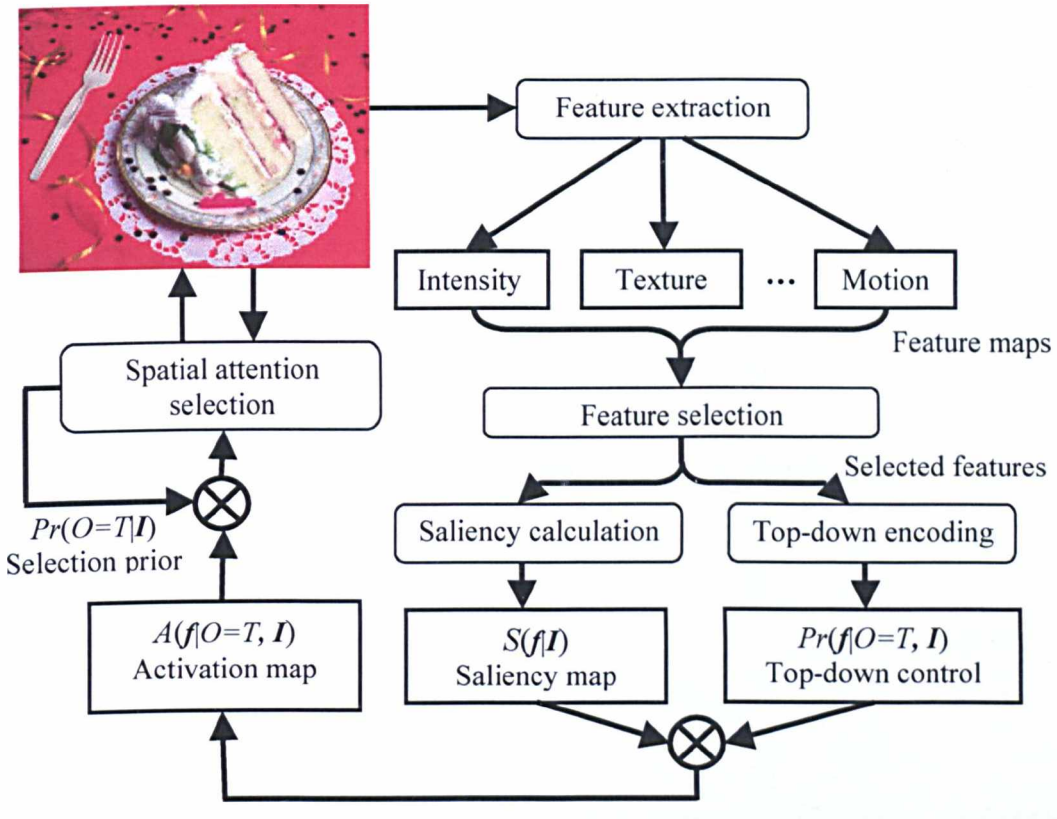



**Figure 7.1** A schematic diagram of the Bayesian modelling of visual attention selection. The input image is decomposed in a parallel fashion into visual features, among which relevant features are selected. Saliency map and top-down control are then computed in the selected feature space. Saliency map drives visual attention to regions with rare features, whereas top-down activation directs visual attention towards regions sharing similar features of the target. The bottom-up saliency is modulated with the top-down bias into a coherent activation map, based on which the input image is sequentially attended in an order of decreasing activation.

### 7.1.2 Visual Feature Selection

An efficient strategy to locate a target is to search for objects having features with larger $Pr(O=T \mid \boldsymbol{f})$, as demonstrated in (7.1). Therefore, irrelevant features can be defined as those that do not affect the activation of objects. Accordingly, the definition of irrelevance can be achieved by the following:

***Definition* 7.1:** Given a set of features $\boldsymbol{f}^{(1)}=\{f_i^{(1)}, 1\le i\le N_1\}$, a feature set $\boldsymbol{f}^{(2)}=\{f_i^{(2)}, 1\le i\le N_2\}$ is irrelevant to attention selection if

$$Pr\left(O=T \mid \boldsymbol{f}^{(1)}, \boldsymbol{I}\right)= Pr\left(O=T \mid \boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)}, \boldsymbol{I}\right), \text{ whenever } Pr\left(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid \boldsymbol{I}\right)\neq 0. \quad (7.5)$$

By using (7.1), we have another equivalent definition:

***Definition* 7.2:** Given a set of features $\boldsymbol{f}^{(1)}=\{f_i^{(1)}, 1\le i\le N_1\}$, $\boldsymbol{f}^{(2)}=\{f_i^{(2)}, 1\le i\le N_2\}$ is irrelevant to attention selection if

$$A\left(\boldsymbol{f}^{(1)} \mid O=T, \boldsymbol{I}\right)= A\left(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid O=T, \boldsymbol{I}\right), \text{ whenever } Pr\left(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid \boldsymbol{I}\right)\neq 0. \quad (7.6)$$

where $A(\boldsymbol{f}^{(1)} \mid O=T, \boldsymbol{I})$ and $A(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid O=T, \boldsymbol{I})$ are defined in (7.3).

In order to select relevant features based on Definition 7.2, we define

$$\begin{cases} P_h = \displaystyle\int_{A(\boldsymbol{f} \mid O=T, \boldsymbol{I})>\beta} Pr(\boldsymbol{f} \mid O=T, \boldsymbol{I})\,d\boldsymbol{f} \\ P_f = \displaystyle\int_{A(\boldsymbol{f} \mid O=T, \boldsymbol{I})>\beta} Pr(\boldsymbol{f} \mid \boldsymbol{I})\,d\boldsymbol{f} \end{cases} \quad (7.7)$$

where $\beta$ is a threshold. By changing $\beta$ from $\infty$ to 0, we can obtain the corresponding ROC curve $ROC(\boldsymbol{f} \mid O=T, \boldsymbol{I})$ and the area under the ROC curve $AUC(\boldsymbol{f} \mid O=T, \boldsymbol{I})$ [197, 198]. By using the same procedure as illustrated in Section 6.2, we obtain an equivalent definition of feature irrelevance, based on which an approximated algorithm can be designed for backward and forward feature selection.

***Definition* 7.3:** Given a set of features $\boldsymbol{f}^{(1)}=\{f_i^{(1)},1\le i\le N_1\}$, $\boldsymbol{f}^{(2)}=\{f_i^{(2)},1\le i\le N_2\}$ is irrelevant to attention selection if

$$AUC\left(\boldsymbol{f}^{(1)} \mid O=T,\boldsymbol{I}\right)=AUC\left(\boldsymbol{f}^{(1)},\boldsymbol{f}^{(2)} \mid O=T,\boldsymbol{I}\right) \tag{7.8}$$

where $AUC(\boldsymbol{f}^{(1)} \mid O=T, \boldsymbol{I})$ and $AUC(\boldsymbol{f}^{(1)},\boldsymbol{f}^{(2)} \mid O=T, \boldsymbol{I})$ are calculated according to (7.7).

The approximation algorithm of backward elimination for selecting $K$ features is as follows:

(a) Let $\boldsymbol{f}^{(k)}$ be the full feature set and $k$ be the size of the full feature set.

(b) Calculate the discriminability differential matrix $M(f_i,f_j)$; $f_i\in\boldsymbol{f}^{(k)}$, $f_j\in\boldsymbol{f}^{(k)}$, $f_i\neq f_j$.

$$M\left(f_i,f_j\right)=AUC\left(\left\{f_i,f_j\right\} \mid O=T,\boldsymbol{I}\right)-AUC\left(\left\{f_j\right\} \mid O=T,\boldsymbol{I}\right) \tag{7.9}$$

(c) If $k=K$, output $\boldsymbol{f}^{(k)}$.

(d) For $f_i\in\boldsymbol{f}^{(k)}$ $(i=1,k)$

- select $k_s$ features from $\boldsymbol{f}^{(k)}$ to construct a feature subset $\boldsymbol{S}^{(ki)}$. The criterion of the selection is to find feature $f_j$, for which $M(f_i,f_j)$ is the smallest, where $f_j\in\boldsymbol{f}^{(k)}$, $f_j\neq f_i$.
- calculate $D_{AUC}$.

$$D_{AUC}\left(f_i\right)=AUC\left(\boldsymbol{S}^{(ki)}+\left\{f_i\right\} \mid O=T,\boldsymbol{I}\right)-AUC\left(\boldsymbol{S}^{(ki)} \mid O=T,\boldsymbol{I}\right) \tag{7.10}$$

(e) Select feature $f_d$ with the smallest $D_{AUC}$.

(f) $\boldsymbol{f}^{(k)}=\boldsymbol{f}^{(k)}-\{f_d\}$; $k=k-1$; go to (c).

In this algorithm, only a small set of the most informative features is used for computing the change of the AUC when a feature is eliminated. It is evident that the above is similar to that described in Section 6.2.3.

### 7.1.3 Feature Extraction and Probability Density Estimation

The first processing step in the computational model shown in Figure 7.1 is to extract visual features. Computational algorithms have been designed to imitate early visual

feature extraction. For example, Gabor-Wavelets [116] simulates the response of orientation-selective neurons, and the Difference-Of-Gaussian (Mexican hat) filter [115] is implemented to simulate the response of intensity-centre-surround contrast neurons [108]. The 'texture of texture' features [14], proposed by DeBonet *et al.*, are a natural extension of the simple features, including oriented edges and bars, in a recursive way. As illustrated in Section 3.2.3, this method provides a systematic approach for extracting a large number of features to express the causal structures of an image. By using a down-sampling operation, it is able to express the multi-scale representation of image contents. In this study, we use the 'texture of texture' features to demonstrate how the proposed framework works.

The performance of feature selection and saliency calculation relies on the accurate estimation of the likelihood and prior probability densities. One of the simplest and most widely used density estimators is feature histogram, which has been applied in the computation of the AUC to select relevant features in the framework. In top-down encoding and bottom-up saliency calculation, the Gaussian Mixture [217] can be used to model the probability densities for selected continuous features.

Suppose the number of components $g_i(\boldsymbol{f}/\theta_i)$ in a Mixture of Gaussian is $G$, where $\theta_i$ denotes the parameters of $g_i$ and $\boldsymbol{f}$ is a feature point in the selected feature space. The mixture of $pg(\boldsymbol{f}/c,\theta)$ can be expressed as

$$pg(\boldsymbol{f} \mid c,\theta) = \sum_{i=1}^{G} c_i g_i(\boldsymbol{f} \mid \theta_i) \tag{7.11}$$

where

$$\sum_{i=1}^{G} c_i = 1 \text{ and } \begin{cases} c = (c_1, c_2, \ldots c_G) \\ \theta = (\theta_1, \theta_2, \ldots \theta_G) \end{cases}. \tag{7.12}$$

For a Gaussian component,

$$\begin{cases} g_i(\boldsymbol{f} \mid \theta_i) = \dfrac{1}{\left(\sqrt{2\pi}\right)^N \sqrt{|\Sigma_i|}} exp\left( -\dfrac{1}{2}(\boldsymbol{f} - \mu_i)^T (\Sigma_i)^{-1} (\boldsymbol{f} - \mu_i) \right) \\ \theta_i = (\mu_i, \Sigma_i) \end{cases} \tag{7.13}$$

where $\mu_i$ is the mean of the Gaussian component and $\Sigma_i$ is the covariance matrix.

*freq(f)* is used to denote the frequency of feature vector *f*. The parameters $c$ and $\theta$ of the Gaussian Mixture can be determined by the following recurrent equations, which can be derived from the Expectation-Maximisation (EM) algorithm. In the general form of the EM algorithm [218], an observed data $x$ is considered to be an incomplete data of the complete data $y$, where the many-to-one map from $y$ to $x$ is assumed to be known. The EM algorithm is used to find the local maximum of likelihood $\sum_x[\log p(x \mid \theta)]$, where $\theta$ is the parameter that needs to be identified. The algorithm starts from an initial solution $\theta^{(0)}$, and it develops solutions $\theta^{(1)}$, $\theta^{(2)}$, ..., $\theta^{(t)}$ iteratively. Each iteration step consists of an expectation step (E-step) and a maximisation step (M-step).

The EM algorithm used can be summarised as following [219]:

1. Let $\theta^{(0)}$ be an initial solution.
2. Repeat the following E-step and M-step, for t=0,1,2, ...,
   (a) E-step:
   Calculate the expectation value of log-likelihood of the complete data, conditioned by observed samples and the current solution $\theta^{(t)}$:

$$Q(\theta \mid \theta^{(t)}) = \sum_x \left\{E\left(\log p(y \mid \theta) \mid x, \theta^{(t)}\right)\right\} \tag{7.14}$$

   where the summation is taken over all samples.
   (b) M-step
   Let $\theta^{(t+1)}$ be such $\theta$that maximizes $Q(\theta \mid \theta^{(t)})$.

By applying the EM algorithm to estimate the parameters in (7.11) and (7.13), we have

$$p_i(\boldsymbol{f}) = \frac{c_i^{old} g_i\left(\boldsymbol{f} \mid \mu_i^{old}, \Sigma_i^{old}\right)}{\sum_{k=1}^{G} c_k^{old} g_k\left(\boldsymbol{f} \mid \mu_k^{old}, \Sigma_k^{old}\right)}$$

$$\begin{cases} c_i^{new} = \dfrac{\sum_{\boldsymbol{f}} freq(\boldsymbol{f}) p_i(\boldsymbol{f})}{\sum_{\boldsymbol{f}} freq(\boldsymbol{f})} \\ \mu_i^{new} = \dfrac{\sum_{\boldsymbol{f}} freq(\boldsymbol{f}) p_i(\boldsymbol{f}) \boldsymbol{f}}{\sum_{\boldsymbol{f}} freq(\boldsymbol{f}) p_i(\boldsymbol{f})} \\ \Sigma_i^{new} = \dfrac{\sum_{\boldsymbol{f}} freq(\boldsymbol{f}) p_i(\boldsymbol{f}) \left[\left(\boldsymbol{f} - \mu_i^{old}\right)^T \left(\boldsymbol{f} - \mu_i^{old}\right)\right]}{\sum_{\boldsymbol{f}} freq(\boldsymbol{f}) p_i(\boldsymbol{f})} \end{cases} \tag{7.15}$$

A suitable number of the components is often decided empirically according to the number of the distinct clusters in the feature space. It can also be determined by an automatic process as described in [220].

If the conditional independence between features is assumed, the activation calculation can be simplified as

$$A(\boldsymbol{f} \mid O = T, \boldsymbol{I}) = \frac{Pr(\boldsymbol{f} \mid O = T)}{Pr(\boldsymbol{f} \mid \boldsymbol{I})} \approx \frac{\prod_{i=1}^{N} Pr(f_i \mid O = T)}{\prod_{i=1}^{N} Pr(f_i \mid \boldsymbol{I})} \tag{7.16}$$

$$= \prod_{i=1}^{N} \left( \frac{Pr(f_i \mid O = T)}{Pr(f_i \mid \boldsymbol{I})} \right)$$

where $\boldsymbol{f}=\{f_1, f_2, \ldots, f_N\}$.

The above approximation avoids the estimation of probability densities in a multi-dimensional feature space. This can reduce the computational cost and work well in many cases when the independence assumption is not severely violated [207].

### 7.1.4 Scan Path Analysis for Modelling Attention Selection

Studies on eye movements indicate that information is selected actively during visual search and patterns of generated scan paths are affected by our knowledge, interests and expectations of the scene [6]. In Chapter 4, we have illustrated the framework of hot spot detection for capturing the intrinsic visual strategies of different observers [187]. The scan path is first projected into feature space, which is then followed by calculating the fixation density function to eliminate the scene-dependent projection bias. In this section, we demonstrate its usability in modelling attention selection when the explicit domain knowledge about the target features is not available.

As visual search theory reveals, saccadic eye movements tend to fall on items sharing common features with the target [30]. It is possible to estimate the top-down factor $Pr(\boldsymbol{f} \mid O=T, \boldsymbol{I})$ by using the feature coverage of a scan path generated from eye movements. The fixation density function, shown in Equation (4.6), of a scan path is calculated by normalising feature coverage of the scan path with the scene-dependent feature distribution, which is derived from a simulated scan path that covers the entire image through a uniform sampling procedure without feature preference. As a result, the activation of feature $\boldsymbol{f}$, $A(\boldsymbol{f} \mid O=T, \boldsymbol{I})$, can be approximated by its corresponding fixation density function $\Gamma(\boldsymbol{f})$, as defined in Equation (4.6). Given a scan path $\xi=<x_i, y_i, t_i>$ $(1 \le i \le N)$, where $x_i$ and $y_i$ represent the spatial location of each fixation, $t_i$ represents the corresponding dwell duration and $N$ is the number of fixations, we have

$$\begin{cases} Pr(\boldsymbol{f} \mid O=T, \boldsymbol{I}) \propto T(\boldsymbol{f}) = \sum_{(x_i,y_i)\in\xi} \left( \sum_{(x,y)\in\Omega(x_i,y_i)} t_i P_{\boldsymbol{f}}^{i}(x,y) \right) \\ Pr(\boldsymbol{f} \mid \boldsymbol{I}) \propto T_0(\boldsymbol{f}) = \sum_{(x_i,y_i)\in image} \left( \sum_{(x,y)\in\Omega(x_i,y_i)} P_{\boldsymbol{f}}^{i}(x,y) \right) \end{cases} \tag{7.17}$$

and

$$A(\boldsymbol{f} \mid O=T, \boldsymbol{I}) \approx \Gamma(\boldsymbol{f}) = \frac{T(\boldsymbol{f})/T_0(\boldsymbol{f})}{\sum_{\boldsymbol{f}} T(\boldsymbol{f})/T_0(\boldsymbol{f})} \tag{7.18}$$

where $\Omega(x_i,y_i)$ is the foveal field of fixation $<x_i, y_i>$; $P_f^I(x,y)$ is a Gaussian function centred at $<x_i, y_i>$, which is assigned to the projection $<x_i, y_i> \rightarrow \boldsymbol{f}$, denoting the probability of feature $\boldsymbol{f}$ being targeted at fixation $<x_i, y_i>$.

As explained previously, the fixation density function is closely related to the background of the visual scene and its accuracy relies on the calibration of the eye-tracking system. In addition, some scan paths exhibit systematic patterns, such as left-to-right and top-to-bottom scanning, although they have no functional relevance to the task [89]. Furthermore, fixations can also be targeted to irrelevant positions. For instance, fixations can be located at the weighted centre of potential targets, as depicted by the 'centre-of-gravity' effect [122]. Nevertheless, this approximation allows us to simulate human visual search behaviours by taking into account factors that are subconsciously applied by observers during visual recognition.

### 7.1.5 Spatial Attention Selection

One of the common strategies for spatial attention selection is the 'winner-take-all' strategy [67, 108, 211], which selects the object with the maximum response (saliency or activation) for the next fixation point. Instead of using maximum value in the saliency map to guide attention, the FOVEA system [13] applies a probabilistic selection process to make the choice of the selection less rigid. In FOVEA, the principle of inhibition of return is implemented by using the local suppression criterion. The fixation point is weighted more strongly in the vicinity of the current fixation by the proximity criterion. In the proposed framework, we use the following criteria for spatial attention selection.

- *The principle of inhibition of return*: the principle is implemented by decreasing the prior probability $Pr(O=T \mid \boldsymbol{I})$ of an object after it has been attended. After one fixation centred at $(x_i, y_i)$, the activation of an object $O$ in the image is multiplied by the following factor:

$$D_i(O) = 1 - exp\left( -\frac{(\theta_i(O))^2}{2(\lambda\theta_0)^2} \right) \tag{7.19}$$

The above exponential function approximates the function of visual acuity at various eccentricities [35], where $\theta_i(O)$ is the visual angle from $(x_i, y_i)$ to the location of object $O$, $\lambda$ is a scaling factor, and $\theta_0$ is the radius of the foveal field. In this study, we set $\theta_0$=1.5° and $\lambda$=0.6.

- *Fixation selection*: similar to the FOVEA system [13], we use a probabilistic process to select the next fixation point. A random value $\eta$ is determined over the activation map $A(\boldsymbol{f} \mid O=T, \boldsymbol{I})$, as governed by a modified exponential density:

$$f(x) = k \cdot exp\left(-\alpha\left(A_{max} - x\right)\right),\ x \in [0, A_{max}] \tag{7.20}$$

where $A_{max}$ is the maximum value of the activation map $A(\boldsymbol{f} \mid O=T, \boldsymbol{I})$ and $k$ is a scaling factor to make the integral of $f(x)$ equal to 1 and $\alpha$ is empirically set to 1.5. $f(x)$ guarantees that the generated random value $\eta$ is close to $A_{max}$. As a result, a small set of salient objects can be selected and less salient objects can be filtered out by using $\eta$ to perform a thresholding operation. The next fixation point is then determined within the remaining objects by using the proximity rule, *i.e.* the object nearest to the current fixation is finally selected.

# 7.2 Experimental Results

## 7.2.1 Two-Colour Conjunction Search

Figure 7.2 shows an experiment for a two-colour conjunction search. In the experiment, observers were asked to identify the location of the green/yellow block in the shortest time possible. As demonstrated in the hotspot and sequence analysis in Chapters 4 and 5, the primary colour used by the observers to search for the target was green. The features of the image can be defined in a four-dimensional colour space $F=F_b \times F_r \times F_y \times F_g$, where $F_b$, $F_r$, $F_y$ and $F_g$ have two possible values {1, 0}, representing the existence of the corresponding colour. The target in the experiment can be explicitly defined as

$$\begin{cases} Pr(\boldsymbol{f} \mid O=T, \boldsymbol{I}) = 1; \boldsymbol{f} = \{0,0,1,1\} \\ Pr(\boldsymbol{f} \mid O=T, \boldsymbol{I}) = 0; otherwise \end{cases} \tag{7.21}$$

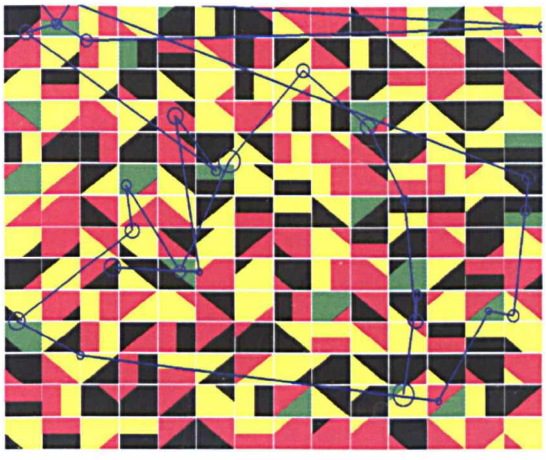


**Figure 7.2** A two-colour conjunction search experiment used for illustrating the key steps involved in the proposed framework. Each small square in the image contains two different colours. The observers were asked to identify the square that contains both green and yellow. Data recording was stopped as soon as the observers had successfully identified the target. The size of the circles represents the dwell time of each fixation whereas the lines between fixations indicate the saccadic eye movements. The primary colour used by the observers for locating the target is green in this experiment.

Figure 7.3 shows the change of the AUC in the backward feature selection with $k_s$=2. The AUC value remains unchanged after the red and black features are eliminated. When yellow is removed, the AUC value decreases slightly from 0.997 to 0.949, which indicates that the discriminability provided by green alone is not much worse than that by black, red, yellow and green together. This result provides a quantitative explanation of the evident green preference by the observers in this experiment. Figure 7.4 shows a simulated scan path by using the method described in Section 7.1.5.

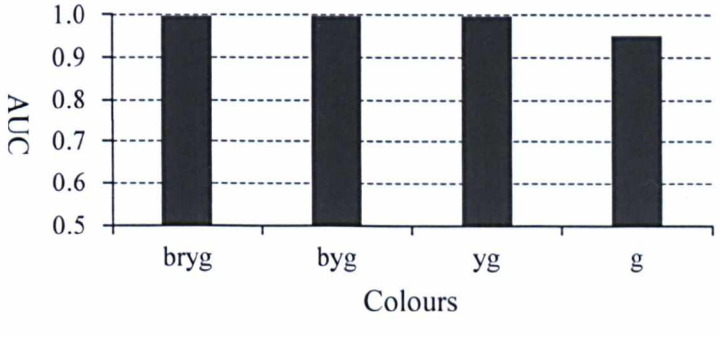



**Figure 7.3** The change of the AUC in the backward feature selection. The AUC value remains unchanged after red and black are eliminated. When yellow is removed, the AUC value decreases slightly from 0.997 to 0.949. 'bryg' denotes that the corresponding AUC is achieved by using the black, red, yellow and green features.

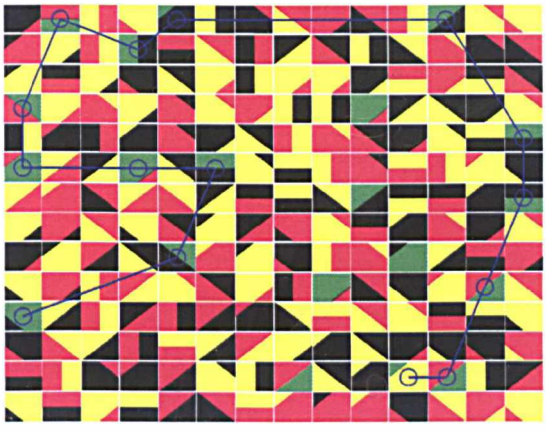


**Figure 7.4** A Simulated scan path. Green is selected by the feature selection algorithm and is used to direct visual attention.

### 7.2.2 Real-World Applications

Figure 7.5 illustrates the HRCT lung image used for this study. 650 feature maps of the 'texture of texture' features described in Section 3.2.3 were extracted to express the texture appearance of the image. A backward selection with $k_s$=3 was applied for feature selection. For the modelling of the foveal vision, a radius of 1.5° for the foveal field was used and the standard deviation of the Gaussian distribution $P_f^i(x,y)$ was 0.25° [187]. Figure 7.6 shows the top-down, bottom-up saliency and activation maps, calculated by following Equations (7.3), (7.17) and (7.18), where the number of selected features was 20, 30 and 40 respectively. The last row highlights regions with $A(\boldsymbol{f} \mid O=T, \boldsymbol{I})>1.0$ to indicate locations where more fixation time is allocated than in a random search. It is obvious that the highlighted regions cover most of the 'black lung' regions, which are regions of interest in this study. Figure 7.7 shows the simulated scan paths based on the derived activation maps. It can be visually identified that the simulated scan paths, derived after feature reduction, and the scan path shown in Figure 7.5 share the same pattern where the 'black lung' regions were frequently attended.

In order to examine the usability of the proposed framework in knowledge gathering, another CT image shown in Figure 7.8 was obtained from the same patient. This CT image is different from the CT image shown in Figure 7.5 in the shape of the parenchyma as well as the spatial distribution of the 'black lung' regions. In a situation like this, traditional spatial analysis cannot be directly applied to model attention selection even after image registration. By using (7.3), (7.17) and (7.18) to extract the top-down information from the scan path shown in Figure 7.5 and to extract the top-down information from the CT image shown in Figure 7.8, we obtain the corresponding top-down map, bottom-up saliency and activation maps (Figure 7.9). Figure 7.10 shows the simulated scan paths based on the derived activation maps. As expected, the 'black lung' regions are the regions of interest for the simulated scan paths, indicating that the acquired feature space representation of visual search patterns can be applied to predict attention selection for other images within the same diagnostic context.

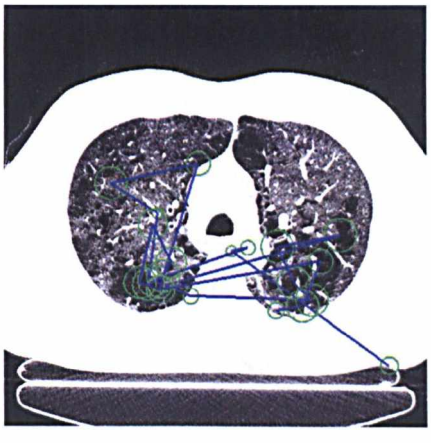


**Figure 7.5** The HRCT lung image of a patient with centrilobular emphysema and the corresponding visual search scan path recorded from an experienced radiologist. The size of the circles represents the dwell time of each fixation and the lines between fixations indicate the saccadic eye movements.

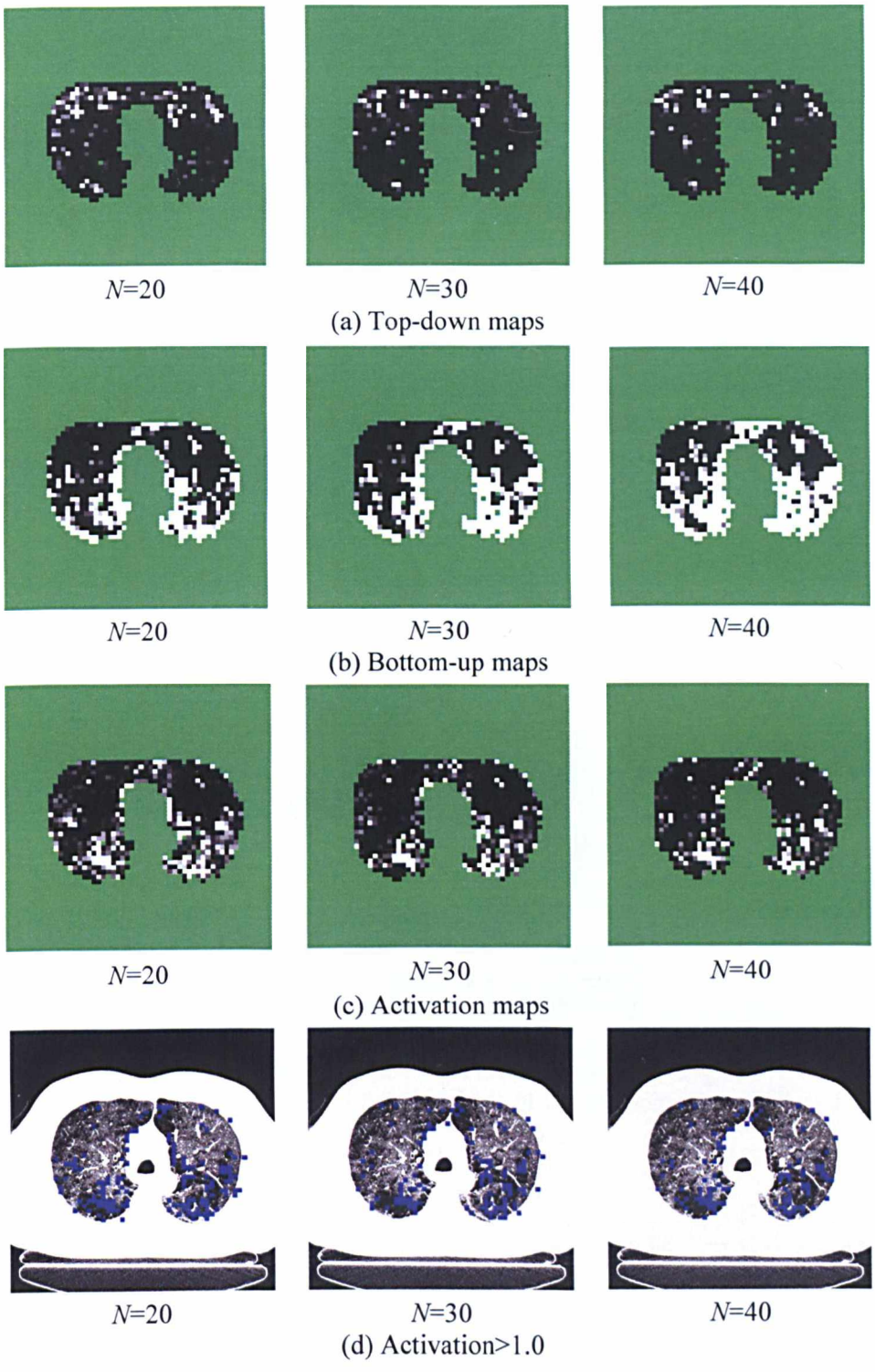
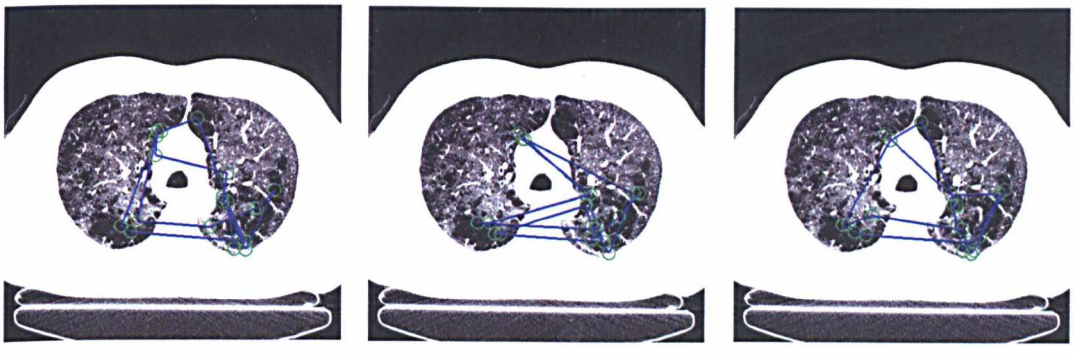



**Figure 7.6** The top-down, bottom-up and activation maps of the CT image shown in Figure 7.5. The number of selected features is 20, 30 and 40 respectively. Brighter grey shades within the lung parenchyma in (a), (b) and (c) signify higher values. Those pixels highlighted in the images of (d) have an activation value greater than 1.0.



*N*=20 *N*=30 *N*=40

**Figure 7.7** The simulated scan paths by using 20, 30 and 40 selected features of the CT image shown in Figure 7.5. The circles represent fixations and the lines between fixations indicate the saccadic eye movements.

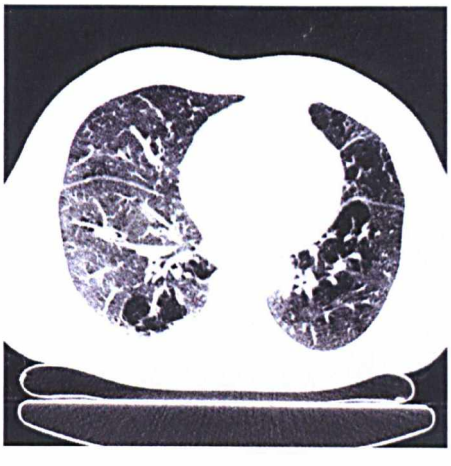


**Figure 7.8** Another CT image was obtained from the same patient of Figure 7.5. It is different from the CT image shown in Figure 7.5 in the shape of the parenchyma as well as the spatial distribution of the 'black lung' regions.

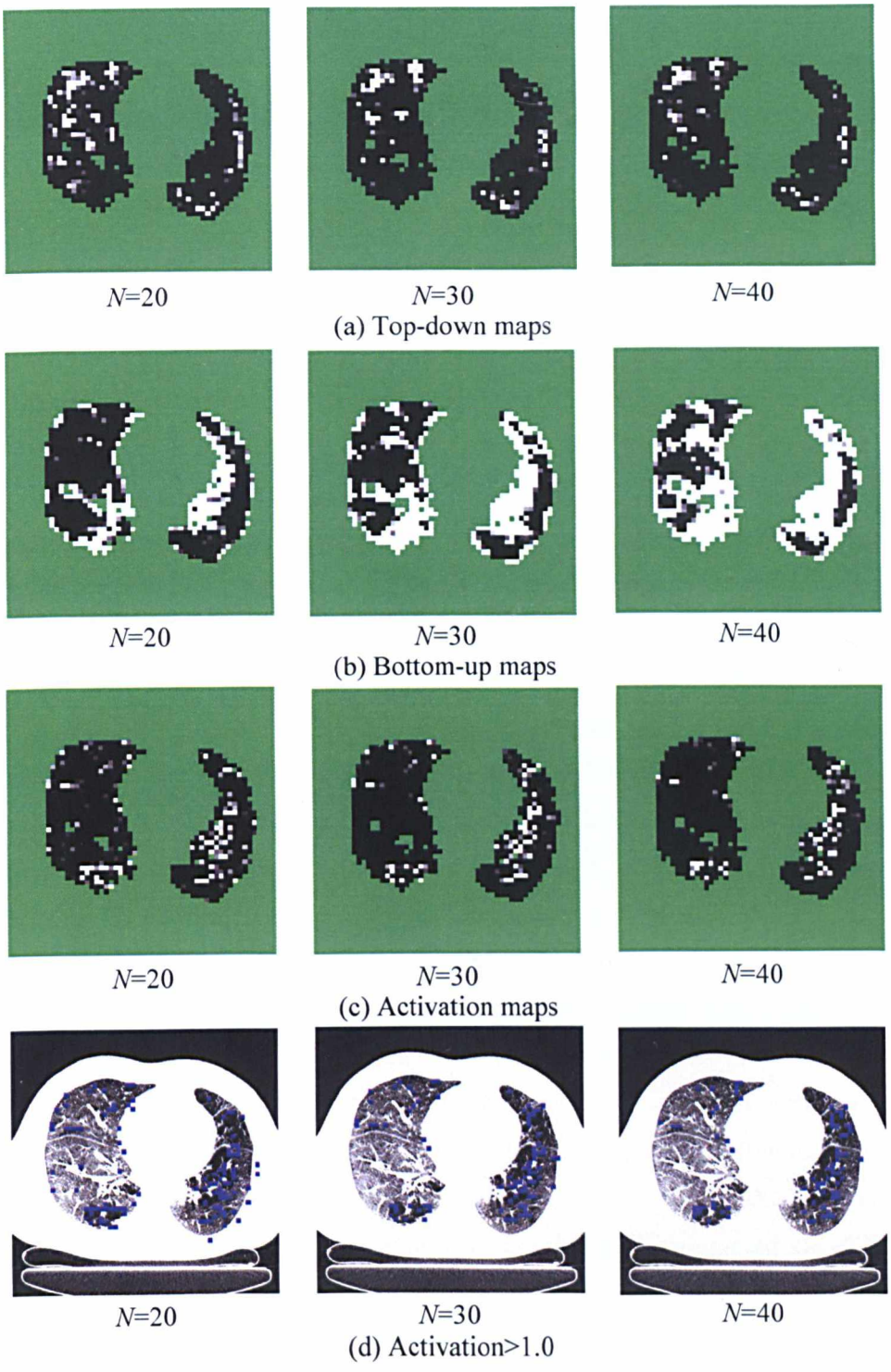
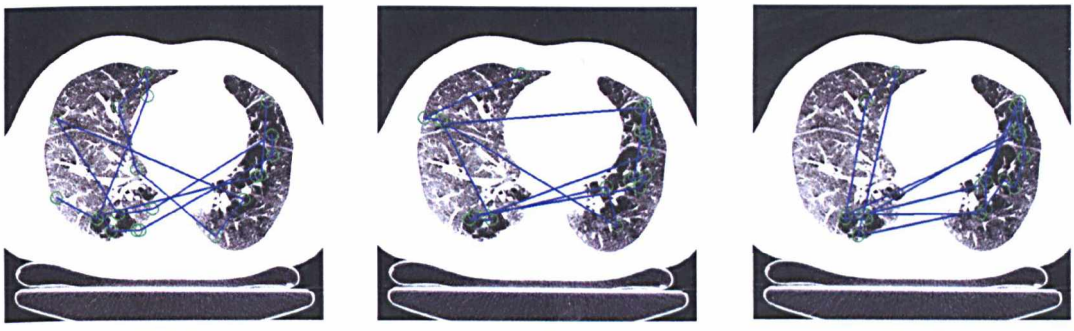



**Figure 7.9** The top-down, bottom-up and activation maps of the CT image shown in Figure 7.8. The number of selected features is 20, 30 and 40 respectively. Brighter grey shades within the lung parenchyma in (a), (b) and (c) signify higher values. Those pixels highlighted in the images of (d) have an activation value greater than 1.0.



*N*=20 *N*=30 *N*=40

**Figure 7.10** The simulated scan paths by using 20, 30 and 40 selected features of the CT image shown in Figure 7.8. The circles represent fixations and the lines between fixations indicate the saccadic eye movements.

The proposed feature space representation of visual search patterns also provides a new possibility for scan path comparison, which has been an important research topic in visual search studies [6, 11]. As demonstrated in the two-colour conjunction search experiment, spatial patterns can be different even when the same search strategy is used. This diversity makes it extremely difficulty to compare scan paths in the spatial domain [211]. Since the activation (or saliency) of a visual feature is determined by its bottom-up and top-down properties and it plays an essential role in the deployment of visual attention [2, 67], the comparison of spatial scan paths can be achieved by studying the corresponding activation maps predicted by the proposed computational model. With this method, we avoid the direct evaluation of spatial fixation patterns. According to the 'winner-take-all' principle [67, 108], visual attention is directed towards locations with higher activity in the activation map. Therefore, the rank of each pixel in the activation map is more important than its exact value. Thus, the nonparametric Spearman rank-order correlation [221], which is based on ranking the two variables and makes no assumption about the distribution of the variables, is suitable for measuring the relationship between two activation maps. The Spearman correlation coefficient is computed by using the following formula:

$$r_s = 1 - \frac{6\sum_{Num} D^2}{Num\left(Num^2 - 1\right)} \tag{7.22}$$

where $Num$ is the number of pixel pairs in the activation maps and $D$ is the rank difference between each pixel pair . Since the down-sample-by-three operation is applied

twice in the feature extraction, the number of degrees of freedom is estimated by (*df=Num*/81-2). In order to convert the activation into a rank order, the activation level of each pixel in the activation map is calculated as follows:

| activation | <1.0/3.0 | 1.0/3.0~1.0/2.0 | 1.0/2.0~1.0/1.25 | 1.0/1.25~1.25 | 1.25~2.0 | 2.0~3.0 | >3.0 |
|---|---|---|---|---|---|---|---|
| level | -3 | -2 | -1 | 0 | 1 | 2 | 3 |

The overall accuracy is also calculated to compare the 'activation>1.0' maps. It is the percentage of the number of the pixels correctly predicted against the total number of the pixels in the 'activation>1.0' maps.

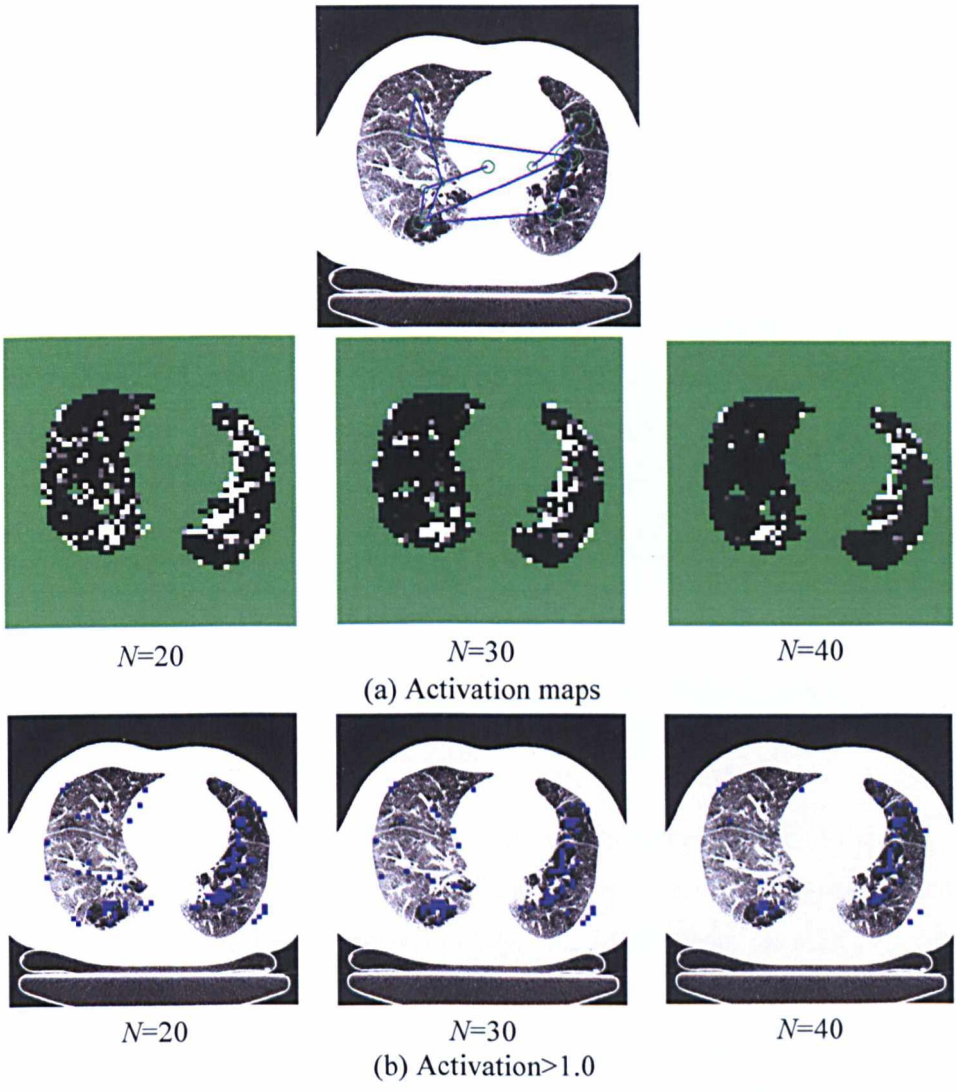



**Figure 7.11** The CT image shown in Figure 7.8 is identical to the above CT image. The superimposed scan path was recorded from the same radiologist of the scan path shown in Figure 7.5. The number of selected features is 20, 30 and 40. Those pixels highlighted in the images of (b) have an activation value greater than 1.0.

The scan path shown in Figure 7.11 was recorded when the same radiologist of the scan path shown in Figure 7.5 was asked to identify common visual features for diagnosis. The displayed activation maps shown in Figure 7.11 are calculated by using (7.3), (7.17) and (7.18). The number of selected features is 20, 30 and 40 respectively. Table 7.1 shows the results of comparing the activation maps shown in Figures 7.9 and 7.11.

**Table 7.1** Results of comparison between the activation maps shown in Figures 7.9 and 7.11.

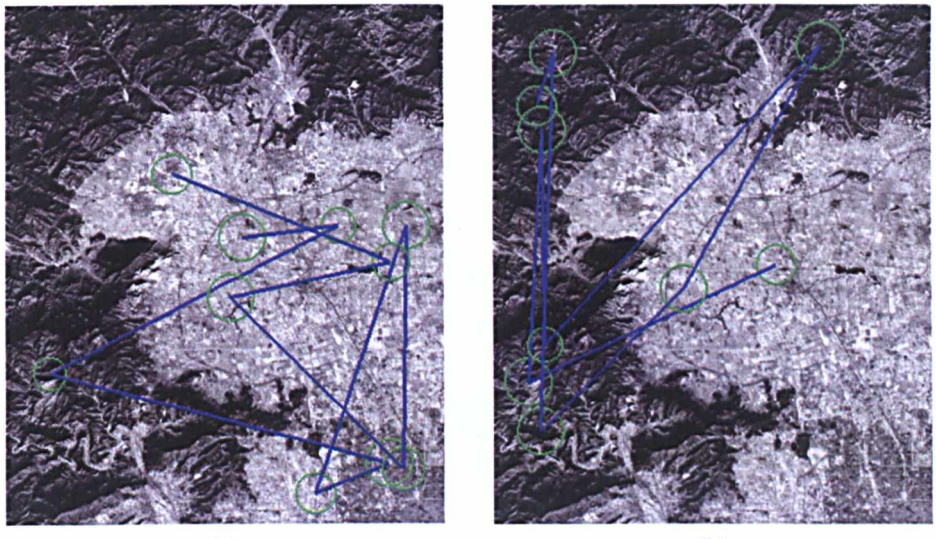
| | $N$=20 | $N$=30 | $N$=40 |
|---|---|---|---|
| Spearman correlation coefficient | 0.534 | 0.692 | 0.757 |
| Two-tailed significance level | $p<0.0001$ | $p<0.0001$ | $p<0.0001$ |
| Overall accuracy | 81.33% | 85.27% | 87.59% |

The low significance levels ($p<0.001$) of the correlation coefficients and large overall accuracies suggest a strong positive correlation between the two activation maps.

The above method was also applied to analyse visual search behaviours for satellite images. Figure 7.12 shows two scan paths superimposed on a satellite image of a part of Beijing city [222]. Scan path (a) was obtained when an observer was asked to search for details of the city, while scan path (b) was obtained when the observer was asked to search for features in the mountain area. The two tasks yielded two different scan paths. Figure 7.13 shows three satellite images of three different parts of Beijing. For feature space analysis, the 650 'texture of texture' feature maps were extracted to express the texture appearance of the images. Equations (7.3), (7.17) and (7.18) were applied to calculate the top-down map, bottom-up saliency and activation maps. Twenty features were selected by the backward selection with $k_s$=3 for each image.

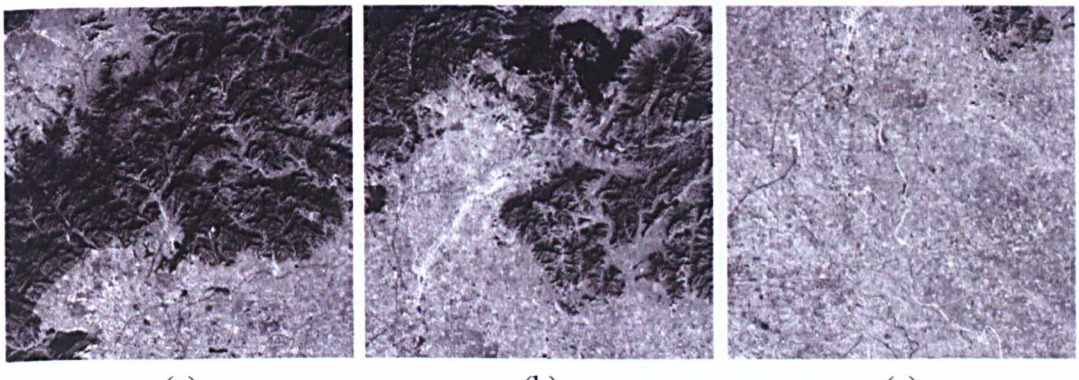
Figures 7.14 and 7.15 show the activation maps and simulated scan paths corresponding to the scan paths shown in Figures 7.12(a) and (b) respectively. For images shown in Figures 7.13(a) and (b), the high activation regions, derived from the scan path shown in Figure 7.12(a) when the observer was asked to search for details of the city, cover the urban areas. In contrast, the high-activation regions, derived from the scan path shown in Figure 7.12(b), cover the mountain areas. It is interesting to see that the mountain area shown in Figure 7.13(c) 'pops out' for both scan paths illustrated in Figures 7.12(a) and (b). Since the mountain area of Figure 7.13(c) occupies a very small proportion of the whole image, it becomes salient even for the scan path shown in Figure 7.12(a). In this

situation, the bottom-up factor plays a dominant role in attention deployment. In order to cross-validate the above results, a set of scan paths were collected when the observers were asked to search for the details of the city and the mountains on the satellite images shown in Figure 7.13. Figures 7.16 and 7.17 show the recorded scan paths and the corresponding activation maps. Tables 7.2 and 7.3 are the results of comparing the activation maps shown in Figures 7.14 and 7.16, and in Figures 7.15 and 7.17 respectively. Once again, the large overall accuracies suggest a reliable positive correlation between the activation maps.



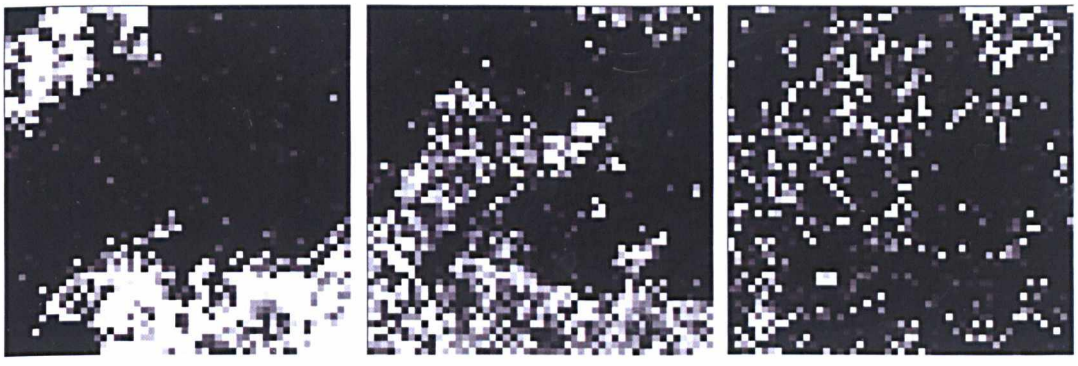
(a) (b)

**Figure 7.12** A satellite image of a part of Beijing city and two superimposed scan paths. Scan path (a) was obtained when an observer was asked to search for the details of the city, while scan path (b) was obtained when the observer was asked to view the mountain appearance.



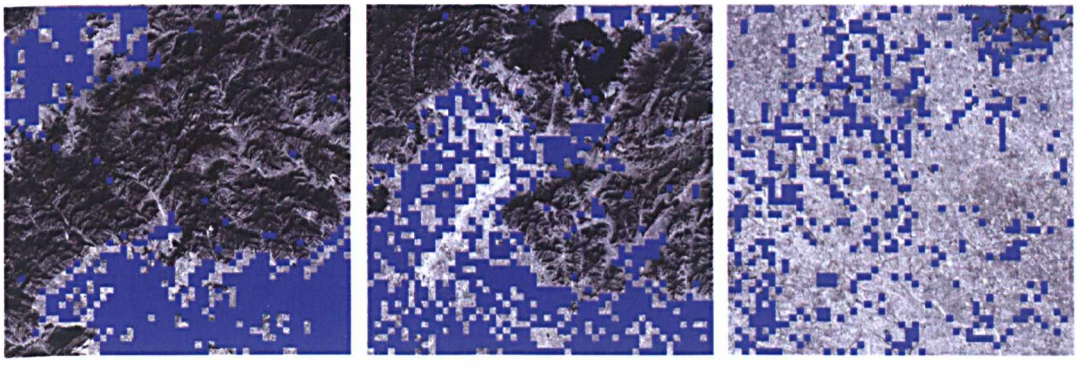
(a) (b) (c)

**Figure 7.13** Three satellite images of three different parts of Beijing city.



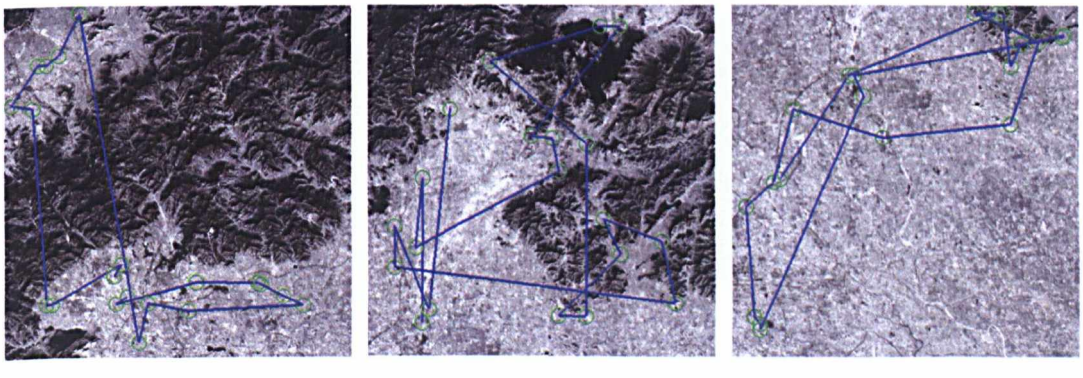
(a) (b) (c)

Activation maps



(a) (b) (c)

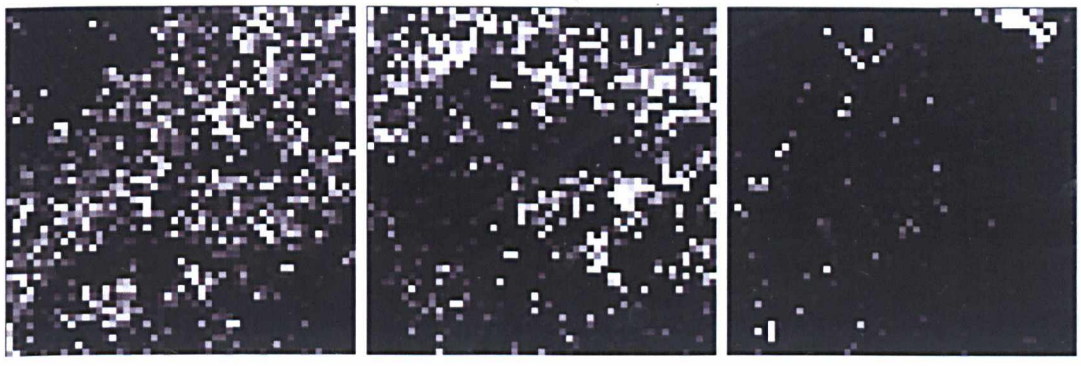
Activation>1.0



(a) (b) (c)

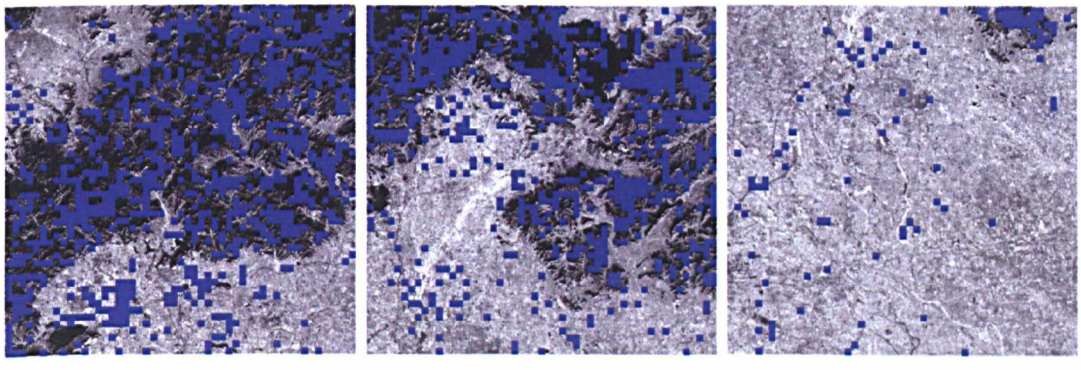
Simulated scan paths

**Figure 7.14** The activation maps and simulated scan paths corresponding to the scan path shown in Figure 7.12(a) for the images shown in Figure 7.13 by using 20 selected features. Brighter grey shades in the activation maps signify higher values. Those pixels highlighted in the images of the second panel have an activation value greater than 1.0.



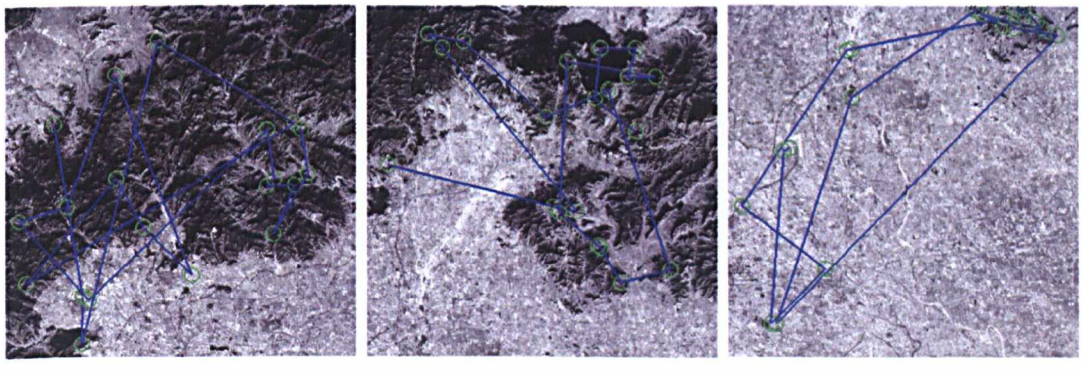
(a) (b) (c)

Activation maps



(a) (b) (c)

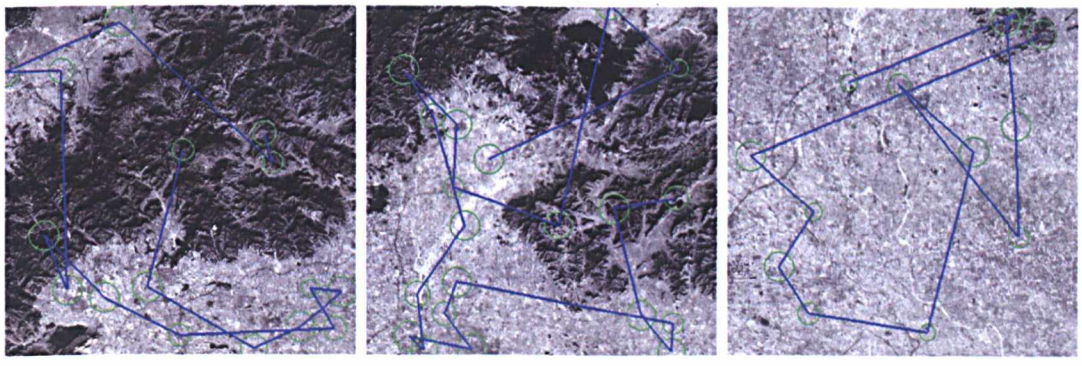
Activation>1.0



(a) (b) (c)

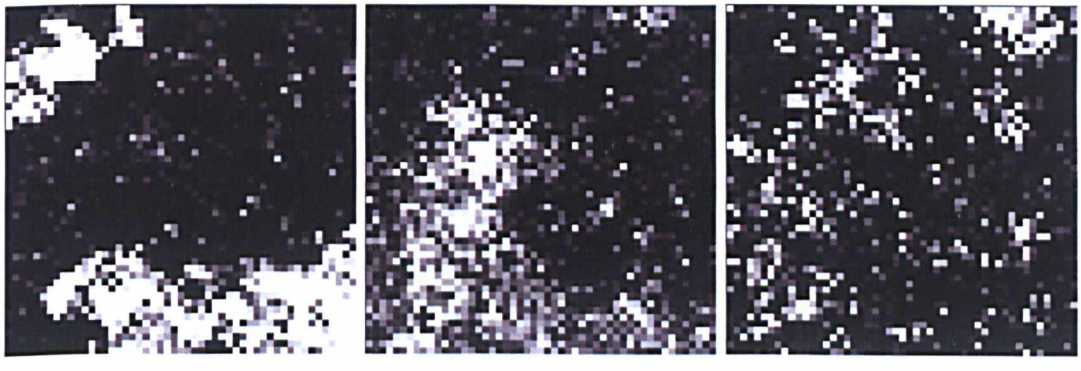
Simulated scan paths

**Figure 7.15** The activation maps and simulated scan paths corresponding to the scan path shown in Figure 7.12(b) for the three images shown in Figure 7.13 by using 20 selected features. Brighter grey shades in the activation maps signify higher values. Those pixels highlighted in the images of the second panel have an activation value greater than 1.0.



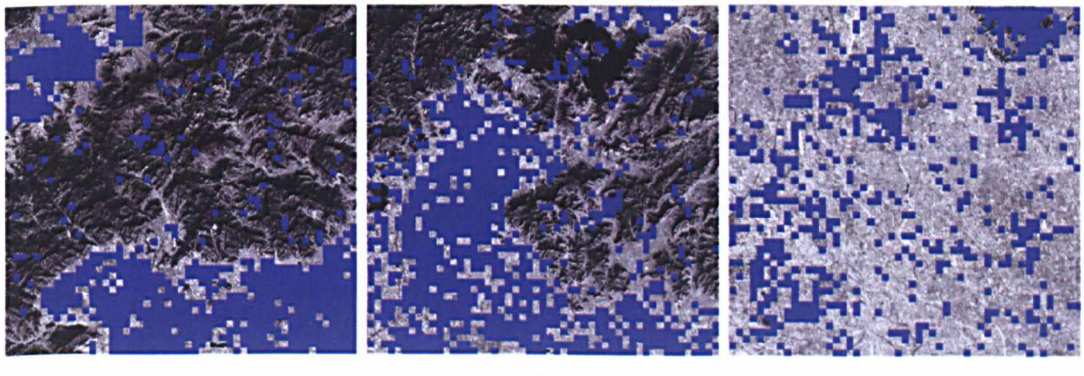
(a) (b) (c)

Satellite images and scan paths



(a) (b) (c)

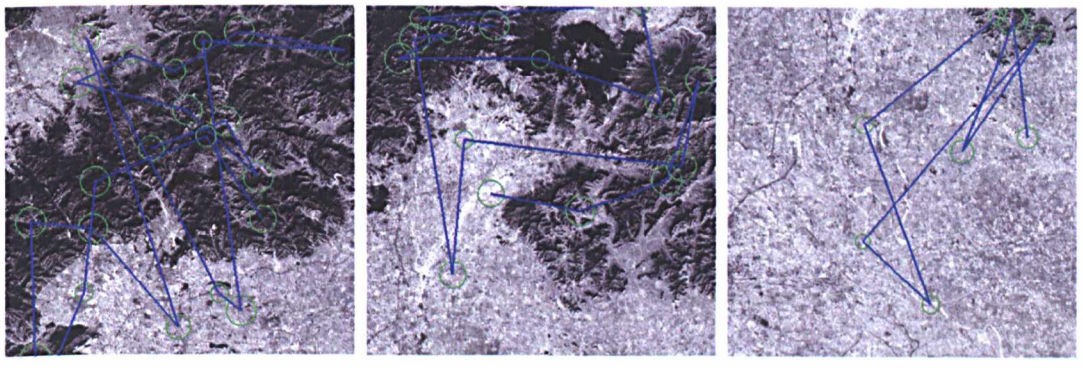
Activation maps



(a) (b) (c)

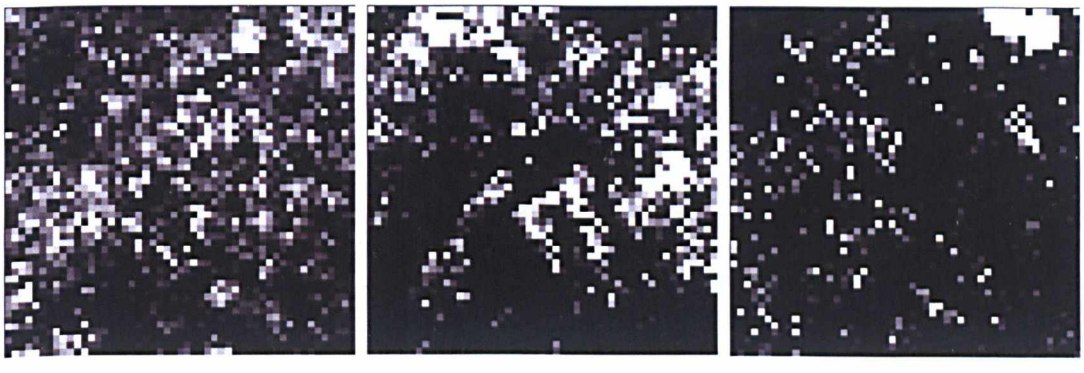
Activation>1.0

**Figure 7.16** The superimposed scan paths were recorded from the same observer of the scan path shown in Figure 7.12 when searching for the details of the city. The number of selected features is 20. Brighter grey shades in the activation maps signify higher values. Those pixels highlighted in the images of the bottom panel have an activation value greater than 1.0.



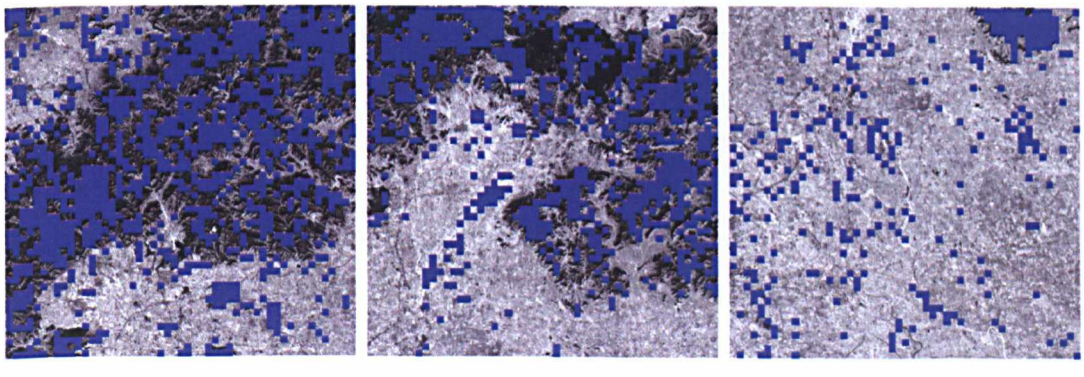
(a) (b) (c)

Satellite images and scan paths



(a) (b) (c)

Activation maps



(a) (b) (c)

Activation>1.0

**Figure 7.17** The superimposed scan paths were recorded from the same observer of the scan path shown in Figure 7.12 when searching for the details of the mountains. The number of selected features is 20. Brighter grey shades in the activation maps signify higher values. Those pixels highlighted in the images of the bottom panel have an activation value greater than 1.0.

**Table 7.2** Results of comparison between the activation maps shown in Figures 7.14 and 7.16.

| | (a) | (b) | (c) |
|---|---|---|---|
| Spearman correlation coefficient | 0.649 | 0.507 | 0.474 |
| Two-tailed significance level | $p<0.0001$ | $p<0.0001$ | $p<0.0001$ |
| Overall accuracy | 80.64% | 70.40% | 77.24% |

**Table 7.3** Results of comparison between the activation maps shown in Figures 7.15 and 7.17.

| | (a) | (b) | (c) |
|---|---|---|---|
| Spearman correlation coefficient | 0.205 | 0.433 | 0.672 |
| Two-tailed significance level | $p<0.0001$ | $p<0.0001$ | $p<0.0001$ |
| Overall accuracy | 65.67% | 73.07% | 89.14% |

## 7.3 Discussion and Summary

In this chapter, we have presented a biologically plausible framework for modelling the deployment of visual attention. Bayesian theory is used to describe the interaction between top-down and bottom-up process. Visual features are extracted in the early stage of the processing, followed by a selection process to evaluate the relevance of extracted features. Instead of using all visual features to construct a feature space for calculating saliency maps, the proposed method uses a subset of relevant features to reduce the complexity of scene analysis. We have demonstrated how to incorporate explicit domain knowledge of human visual search behaviours into the framework. A probabilistic implementation of the 'winner-take-all' principle [67, 108, 211] is applied to the simulation of spatial attention deployment. The primary feature of the proposed model is the selection of relevant visual features. Thus far, methods such as PCA have been employed to model visual information [212, 213]. The interaction between top-down and bottom-up visual information has mainly been used to explain the spatial selection of visual attention. In this study, we have shown that visual feature selection is also a consequence of feature interaction and should be integrated into the modelling process. An optimal algorithm has been designed for feature selection by taking into account both top-down and bottom-up information. Experimental results indicate the strengths of the proposed model in selecting relevant visual features for spatial attention deployment.

A major limitation of the proposed approach is the ignorance of patterns of fixation sequence, which can be crucial in object recognition for establishing the relationships between features and to resolve ambiguities [6]. It should be noted that observers may adopt different visual search strategies in different spatial regions of the image if the feature distributions in these regions are different [187]. It is therefore necessary to consider visual segmentation and spatial analysis. Recent studies also indicate that strategies and visual contextual information about past search experience in similar environments play an important role in the formation of saliency maps [213]. The question of how to incorporate those factors into the framework requires further investigation.

In conclusion, the framework of Bayesian feature selection can be applied to the modelling of visual attention selection and for scan-path analysis in feature space. It can serve as a theoretical basis for explaining the interaction between top-down information, *i.e.* the observer's expectation and prior knowledge, and the bottom-up process, *i.e.* task-independent visual sensory stimulation, in visual perception. By following this framework, the dominant features for visual search are those that can provide high discriminability for distinguishing search targets from the background.

# Chapter 8

# Conclusions and Future Work

## 8.1 Summary of the Thesis

Medical image understanding is primarily a subjective process, and visual feature comparison and cross-examination by an experienced radiologist involve both conscious and subliminal levels of processing. As a result, the understanding of how experts' visual search takes place is difficult to achieve by traditional knowledge-gathering methods. In this work, we have investigated both theoretical and practical issues involved in discovering cognitive visual search strategies for diagnostic decision support systems. Algorithms for feature extraction and scan path acquisition have been described. A framework for hot spot detection, including feature extraction, feature selection, saliency detection, visual pattern integration and back projection has been proposed for capturing intrinsic visual search behaviour of different observers in image understanding. With this framework, we have concentrated on texture features to identify image contents inside the lung parenchyma. The detected hot feature spots can be mapped back to the spatial domain for highlighting regions of interest, which can then be used to determine meaningful transitions in the feature space. The patterns of visual attention selection in response to search tasks are detected by using the signal detection theory. They are represented by hot feature spots (or salient feature values) and the first-order fixation transition matrices. We have demonstrated how to integrate the patterns obtained from different observers to provide an objective assessment of the underlying patterns. A feature selection algorithm based on Bayesian theory was designed to select relevant features for constructing a feature space in which visual search is performed. We have

proposed a Bayesian computational framework to describe the interaction between top-down and bottom-up process, thereby providing a theoretical basis for modelling visual feature selection and attention deployment.

Due to the limited ability of our biological visual system, one of the main characteristics of visual search is selective attention, which plays an important role in many aspects of visual processing. It is well recognised that visual attention is governed by both bottom-up and top-down activities, and can be performed in both spatial domain and feature space. Therefore, most computational models of visual attention take into account both bottom-up information, *i.e.* information about task-independent visual sensory stimulation, and top-down information, *i.e.* task-dependent prior knowledge, interests and expectations. Through this approach, the visual scene is decomposed into semantically related parts, which are then combined with top-down information into a saliency map. This two-stage process is also applied to the proposed framework of hot spot detection and fixation sequence analysis, in which both scan path and visual scene are initially projected into the feature space, followed by the identification of fixation transition patterns and hot spots.

In this thesis, the image features we have used are mainly texture features. Texture information is one of the fundamental primitives in human vision, and the texture appearance of the lung parenchyma provides essential information for the differentiation of various signs of obstructive lung diseases. In this study, local textural description was encapsulated by using the statistical moments of CT densities, co-occurrence matrices, and run-length parameters of neighbourhood pixels. Another set of features included for building a large and informative feature library is the 'texture of texture' features, which are highly selective to the causal structures of images, and are calculated in a recursive way. The main anatomical structures, including the chest wall and pulmonary vessels inside the parenchyma, are removed before texture extraction since they provide no information about the parenchyma textures. An automated segmentation of these anatomical structures is implemented by using defined thresholding operations on CT densities.

For the work reported in this thesis, a DICOM image viewer was implemented to recreate a natural reporting environment, which was integrated with an unobtrusive, high resolution ASL-504 eye tracking system to acquire reliable scan paths of the radiologists.

This platform is able to record not only the observer's eye movements but also the control operations for interactive multi-slice or volumetric CT manipulations.

The feature space representation of a visual scene and its corresponding visual search scan-path is important to the detection of attention selection by analysing saccadic eye movements. To this end, pixels covered by the scan path are projected to the feature space by considering its associated feature vector and fixation dwell time. The foveal field activation is modelled by a Gaussian density function, since the visual acuity drops off dramatically from the centre of focus. Features corresponding to pixels falling within the foveal field of view are assigned a non-zero probability with value inversely proportional to the spatial distance to the centre of the fixation. The total dwell time of a feature is the sum of the dwell time for each fixation, weighted by the Gaussian density.

In Chapter 4, we have stressed the importance of addressing the bias in visual features due to their relative abundance in the scene. The correct identification of attention selection in feature space requires the elimination of the scene-dependent projection difference, which is independent of visual search strategies. In the framework of hot spot detection, the reliable identification of salient features from visual search scan paths relies on the calculation of feature space density obtained by normalising the feature space fixation distribution with the scene-dependent feature distribution. The rationale of this processing can be explained by using the signal detection theory, in which a signal is detected by comparing the noise response to the noise plus signal response. Feature points with a relatively high fixation density value are considered to be salient or 'hot'. Domain knowledge from radiologists can be used to explain the diagnostic significance of each salient feature, which can be back-projected to the spatial domain for highlighting regions that require specific attention. A two-colour conjunction search test has been designed to demonstrate that the spatial patterns of the scan paths are influenced by the cognitive visual search strategies. A medical imaging experiment has also been applied to illustrate the practical value of the framework for knowledge gathering in diagnostic decision support systems.

The hot spot detection method described in this thesis provides an objective way to determine meaningful transition in feature space for fixation sequence analysis. We have used the first-order transition matrices to detect the patterns of fixation sequences. Similar to hot spot detection, spatial fixation sequences are projected into feature space by calculating the weighted frequency of transitions between features, which are compared

with transition patterns obtained from random scanning in order to eliminate the scene-dependent projection bias. Both artificial and medical images were used to validate the method. The experiments on these images indicate that visual search tasks can become difficult for our visual system when comparisons between objects or features are necessary. This signifies the importance of training and effective knowledge extraction in medical image understanding.

To improve the accuracy and stability of hot spot detection and fixation sequence analysis, we have presented a method to extract common patterns embedded in the scan paths obtained from different observers with similar prior knowledge and experience. The integrated visual search patterns represent the common visual search attention and strategies deployed, and we have proved that a simplified detection procedure in hot spot detection is equivalent to the statistical Bernoulli test, in which each scan path of an observer is used individually to test whether a feature is salient or not. The applicability of this method has been illustrated by the two-colour conjunction search test and the medical imaging experiments as explained above.

In Chapter 6, a feature selection scheme has been designed for dimensionality reduction of the feature space. Since hundreds of features are included in the feature library, it is necessary to reduce its dimensionality by only selecting a small set of features for further processing. The selected features should be able to provide sufficient information for distinguishing attended visual targets from the background scene. The proposed feature selection method is based on information theory and Bayesian theory. We used the Kullback-Leibler (*KL*) divergence and the Area under Receiver Operating Characteristic Curve (AUC) to minimise the amount of predictive information loss during feature elimination. With the proposed Bayesian selection framework, we have attempted to reduce the number of selected features whilst maintaining an expected level of the AUC. To avoid the high computational cost for the AUC calculation in a high-dimensional space, we presented a sub-optimum algorithm in which only a small set of the most informative features is used for computing the change of the expected AUC when a feature is eliminated or added during backward/forward selection. The validation results of the algorithm indicate the potential value of the algorithm in eliminating irrelevant features whilst maintaining/improving the accuracy of the classifiers.

In Chapter 7, we have proposed a computational framework for modelling visual feature selection and attention deployment based on Bayesian theory and proposed a biologically

plausible architecture. In this framework, both top-down and bottom-up visual activities are incorporated into a Bayesian framework under the assumption that the human visual system is optimal in selecting informative locations and features. Both visual feature selection and attention deployment are regarded as consequences of the interaction between top-down and bottom-up visual information. We have explored the possibility of incorporating human visual search behaviours detected by eye movement analysis into scan path simulation. The experimental results indicate that the algorithm is able to select the most informative visual features to construct a feature space for the analysis of visual search patterns.

## 8.2 Future Work

Eye movement is an important activity of human vision. Recent findings from brain imaging and neurobiological research provide evidence for the interpretation of visual search behaviours. Techniques such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) have been utilised to uncover the mechanisms of attention selection, visual memory, the time course of visual processing, and object recognition. These techniques provide extensive information about human vision, which could be combined with eye movement tracking for knowledge gathering. Eye movement analysis in feature space is the focus of this study so far. It could be helpful in future studies to investigate the possibility of incorporating other measurements of human vision, for instance brain imaging, into the current framework. The further investigation of the relationship between eye movements and brain activities is also important for understanding cognitive visual search strategies.

It should be noted that the performance of the proposed method could be affected by many factors. In this thesis, we have used a Gaussian density function with fixed parameters in the modelling of the resolution distribution of the fixation field. This is a potential drawback of the study as the resolution distribution is asymmetric and dynamic for human vision. Furthermore, the abilities of the human visual system in assimilating visual information from different visual features are not identical. For instance, we are more adapted to distinguish objects by their respective colours rather than their shapes in certain circumstances. However, these abilities are treated equally in this study. Such inaccuracy can introduce errors in the feature projection of a scan path and background scene, and therefore the performance could deteriorate.

It is also worth noting that, in this thesis, issues related to parallel search and peripheral vision have not been included in the detection of visual search strategies. Psychological evidence indicates that visual attention can be directed covertly to peripheral visual targets without eye movements [44, 45]. In the Global-Focal model proposed by Nodine *et al* [223], for instance, a '*global impression*' is obtained by peripheral vision and a few early fixations. The use of eye-tracking techniques is particularly suited for serial search where detailed visual examinations are required.

Attention selection is an active and dynamic process. The brain is capable of adjusting the configuration of its visual system to adapt to different environments according to the specific demands of different search tasks. Observers may use certain visual search strategies in different spatial regions of an image if the feature distributions in these regions are different. Visual search is often a self-learning process, and the visual search strategy can be adjusted over time in accordance with the accumulated information content. Furthermore, our visual system is able to adapt different search strategies to work efficiently over different spatial scales. There exists psychophysical evidence that the effective visual field is not constant; instead, it is affected by the signal-to-noise ratio of the scene [185]. As a consequence, it is natural to take into account not only image features, but also spatial and temporal scales in the modelling of the visual search strategy. In the framework of hot spot detection, we have provided a preliminary demonstration of the temporal changes of attention selection. The extant literature indicates that linear-scale space theory [128] can be used in feature extraction, image understanding [135] and segmentation. For instance, Lifshitz *et al.* employed a multi-resolution hierarchical technique to acquire a scale-space image representation, in which the lower resolution provides a global view of the image and the higher resolution provides the detail for segmenting images into regions of interest [224]. The main remaining problem is the incorporation of these factors into a coherent framework for modelling visual search behaviours.

Automatic feature construction and extraction is a logical next step in the evolution of scan path analysis. The performance of the proposed feature space analysis is potentially limited by the incompleteness of the feature library. To forgo the requirement of the construction of explicit feature libraries, classic algorithms such as PCA and MDS (Multidimensional Scaling) have been suggested for detecting the intrinsic dimensionality of a visual scene. Instead of using the traditional Euclidian distance, Tenembaum *et al.*

[176] used geodesic distances by measuring the distance between data points in the geometric shapes, which preserves the global geometry of the data; whereas the Local Linear Embedding (LLE) algorithm, proposed by Roweis and Saul [177], tried to preserve the local neighbourhood relationships of the data. These techniques, coupled with local fixation-triggered local feature analysis, can prove to be useful for automatic feature construction and extraction.

In addition to visual search strategies on a static scene, which is the major emphasis of this thesis, the extension of the proposed framework to dynamic scene analysis remains another challenge. Most modern eye tracking systems are able to track eye movements and produce a video image sequence of the same visual scene of the observers by using a head-mounted scene camera [137]. This opens up new possibilities for the visual search community to develop generic platforms to examine visual search behaviours in a natural environment. Extensive effort has been directed towards the analysis of visual search behaviour from video image sequences [225]. However, the results from current studies in this field are preliminary and application-dependent. The considerations of extending the framework reported in this thesis should take the following into account.

(a) *Feature Extraction Library*: The feature library should be extended to describe the following video contents.

- *Motion and change*: One of the major differences between video and static scene data is that the objects and features in a video scene can move and change. In a rapidly changing scene, the motion and change can greatly affect visual search behaviours. Current techniques such as feature tacking [226] and optical flow [227, 228] can be used for motion estimation. In addition, head-tracking equipment [137] becomes necessary in order to determine the motion resulting from head movements.

- *Depth*: It has been reported that depth can be extracted from methods such as optical flow field and binocular vision [227, 229]. This information is important for the perception of 3-D visual structures. By correlating the depth information with eye movements, it is possible to analyse how the depth estimation is performed by the human visual system, which is crucial in fields such laparoscopic surgery [230].

- *Scale*: When an object is approaching or leaving the scene camera, the scale level of the observation changes. The adaptation of search strategies over scales is an important aspect of human visual search. The scale-space theory coupled with feature tracking techniques [128, 226] could be applied to extract scale information from a video image sequence.

(b) *Fixation Analysis*: In this thesis, we have used a sequence of fixations and saccades to characterise a scan path for static scene analysis. In a dynamic situation, however, this pattern may not be sufficient for the modelling of eye movements. In object tracking, for instance, smooth pursuit movements play an important role in maintaining fixation on a smoothly moving target [32]. In order to track a fast moving object, the eye will use a combination of saccades and smooth pursuit movements. Characterisation of a scan path in video data analysis needs further investigation due to the requirement of tracking moving objects,.

(c) *Visual Search Strategy Analysis*: To discover visual search strategies from video image sequences, both intra- and inter-image cross-checking of visual features are important. This demands high computational capacity since computing and storing features for each video frame can be extremely expensive. Furthermore, peripheral vision, which cannot be measured by eye-tracking systems, plays an important role in visual perception in a rapidly changing scene. New methods to measure visual activities and new models to describe the underlying mechanism should be developed in order to tackle these difficulties.

The progress of visual search theory, eye tracking and medical imaging techniques makes it possible to design a generic framework to help resolve the difficulties in knowledge gathering for decision support systems. We have demonstrated that eye movement analysis, together with automated feature extraction, image processing, computer vision, and artificial intelligence, can help in the constructing of a diagnostic decision support system for improving the accuracy and overall reproducibility. The work presented in this thesis provides a specific example of such an application. Future challenges involve the development of techniques to acquire much more detailed psychological and physiological measurements about human vision, and improved methods for modelling human visual search activities.

# Appendix A

# Proofs for Theorems 6.1 and 6.2

In this appendix, we show the proofs for Theorems 6.1 and 6.2.

## *Proof for Theorem* 6.1:

(a) *Proof*: If $L(f^{(1)} \| y \neq a, y=a) = L(f^{(1)}, f^{(2)} \| y \neq a, y=a)$, $ROC(f^{(1)} \| y \neq a, y=a) = ROC(f^{(1)}, f^{(2)} \| y \neq a, y=a)$.

Let $(P_f^{(12)}, P_h^{(12)})$ be a vertex on the edge of the curve $ROC(f^{(1)}, f^{(2)} \| y \neq a, y=a)$. For a given threshold $\beta$, we have

$$
\begin{aligned}
P_h^{(12)} &= \int_{L(f^{(1)}, f^{(2)}) \, y \neq a, y=a) > \beta} Pr(f^{(1)}, f^{(2)} \mid y \neq a) df^{(1)} df^{(2)} \\
&= \int_{L(f^{(1)} \, y \neq a, y=a) > \beta} \left( \int_{-\infty}^{+\infty} Pr(f^{(1)}, f^{(2)} \mid y \neq a) df^{(2)} \right) df^{(1)} \\
&= \int_{L(f^{(1)} \, y \neq a, y=a) > \beta} Pr(f^{(1)} \mid y \neq a) df^{(1)} \\
&= P_h^{(1)}
\end{aligned}
$$

where $(-\infty, +\infty)$ denotes no constraint for variables. Similarly, we have

$$P_f^{(12)} = P_f^{(1)}$$

So, for each vertex on the curve $ROC(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \,||\, y \neq a, y=a)$, there is a corresponding vertex on the curve $ROC(\boldsymbol{f}^{(1)} \,||\, y \neq a, y=a)$, and vice versa.

(b) *Proof*: If $ROC(\boldsymbol{f}^{(1)} \,||\, y \neq a, y=a) = ROC(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \,||\, y \neq a, y=a)$, $L(\boldsymbol{f}^{(1)} \,||\, y \neq a, y=a) = L(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \,||\, y \neq a, y=a)$.

Let $(P_f^{(12)}, P_h^{(12)})$ be a vertex on the edge of the convex hull $ROC(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \,||\, y \neq a, y=a)$ and the corresponding threshold of the likelihood ratio be $\beta^{(12)}$. In the space $S^{(12)}$ constructed by features of $\boldsymbol{f}^{(1)}$ and $\boldsymbol{f}^{(2)}$, $\Omega^{(12)}$ denotes the regions where $L(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \,||\, y \neq a, y=a) > \beta^{(12)}$ holds. We have

$$\begin{cases} P_h^{(12)} = \displaystyle\int_{L(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)}\, y \neq a, y=a) > \beta^{(12)}} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq a) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} = \int_{\Omega^{(12)}} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq a) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} \\ P_f^{(12)} = \displaystyle\int_{L(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)}\, y \neq a, y=a) > \beta^{(12)}} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y = a) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} = \int_{\Omega^{(12)}} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y = a) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} \end{cases}$$

Since $ROC(\boldsymbol{f}^{(1)} \,||\, y \neq a, y=a) = ROC(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \,||\, y \neq a, y=a)$, there exists a corresponding vertex point $(P_f^{(1)}, P_h^{(1)})$ on the edge of the convex hull $ROC(\boldsymbol{f}^{(1)} \,||\, y \neq a, y=a)$, and

$$\begin{cases} P_f^{(1)} = P_f^{(12)} \\ P_h^{(1)} = P_h^{(12)} \end{cases}$$

Let threshold $\beta^{(1)}$ be the corresponding threshold of the likelihood ratio. In the space $S^{(1)}$ constructed by features of $\boldsymbol{f}^{(1)}$, $\Omega^{(1)}$ denotes the regions where $L(\boldsymbol{f}^{(1)} \,||\, y \neq a, y=a) > \beta^{(1)}$ holds. We have

$$\begin{cases} P_h^{(1)} = \displaystyle\int_{L(\boldsymbol{f}^{(1)}\, y \neq a, y=a) > \beta^{(1)}} Pr(\boldsymbol{f}^{(1)} \mid y \neq a) d\boldsymbol{f}^{(1)} = \int_{\Omega^{(1)}} Pr(\boldsymbol{f}^{(1)} \mid y \neq a) d\boldsymbol{f}^{(1)} \\ P_f^{(1)} = \displaystyle\int_{L(\boldsymbol{f}^{(1)}\, y \neq a, y=a) > \beta^{(1)}} Pr(\boldsymbol{f}^{(1)} \mid y = a) d\boldsymbol{f}^{(1)} = \int_{\Omega^{(1)}} Pr(\boldsymbol{f}^{(1)} \mid y = a) d\boldsymbol{f}^{(1)} \end{cases}$$

By expanding $\Omega^{(1)}$ into the space $S^{(12)}$, we have

$$\Omega^{(1)+(2)} = \left\{ \left( \boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \right) : \boldsymbol{f}^{(1)} \in \Omega^{(1)}, \boldsymbol{f}^{(2)} \in (-\infty, +\infty) \right\}$$

Once again, $(-\infty, +\infty)$ denotes no constraint for variables. We have

$$\begin{cases} \int_{\Omega^{(1)+(2)}} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq \boldsymbol{a}) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} = \int_{\Omega^{(1)}} \left( \int_{-\infty}^{+\infty} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq \boldsymbol{a}) d\boldsymbol{f}^{(2)} \right) d\boldsymbol{f}^{(1)} = \int_{\Omega^{(1)}} Pr(\boldsymbol{f}^{(1)} \mid y \neq \boldsymbol{a}) d\boldsymbol{f}^{(1)} = P_h^{(1)} \\ \int_{\Omega^{(1)+(2)}} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y = \boldsymbol{a}) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} = \int_{\Omega^{(1)}} \left( \int_{-\infty}^{+\infty} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y = \boldsymbol{a}) d\boldsymbol{f}^{(2)} \right) d\boldsymbol{f}^{(1)} = \int_{\Omega^{(1)}} Pr(\boldsymbol{f}^{(1)} \mid y = \boldsymbol{a}) d\boldsymbol{f}^{(1)} = P_f^{(1)} \end{cases}$$

Now, we prove $\Omega^{(1)+(2)}=\Omega^{(12)}$. Since a point $\boldsymbol{f}$ with $Pr(\boldsymbol{f} \mid y\neq\boldsymbol{a})=0$ and $Pr(\boldsymbol{f} \mid y=\boldsymbol{a})=0$ has no meaning in *ROC* calculation, we can eliminate all the points at which both likelihood distributions are zero. Therefore, we assume that at least one of the likelihood distributions is not equal to zero at all the points in $\Omega^{(1)+(2)}$ and $\Omega^{(12)}$.

We define

$$\Psi = \Omega^{(12)} \cap \Omega^{(1)+(2)}$$
$$\Pi 1 = \Omega^{(12)} - \Psi$$
$$\Pi 2 = \Omega^{(1)+(2)} - \Psi$$

Therefore,

$$\begin{aligned} P_h^{(12)} &= \int_{\Omega^{(12)}} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq \boldsymbol{a}) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} \\ &= \int_{\Pi 1} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq \boldsymbol{a}) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} + \int_{\Psi} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq \boldsymbol{a}) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} \\ &= \left( \int_{\Pi 1} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq \boldsymbol{a}) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} - \int_{\Pi 2} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq \boldsymbol{a}) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} \right) + \int_{\Omega^{(1)+(2)}} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq \boldsymbol{a}) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} \\ &= \left( \int_{\Pi 1} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq \boldsymbol{a}) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} - \int_{\Pi 2} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq \boldsymbol{a}) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} \right) + P_h^{(1)} \end{aligned}$$

Similarly,

$$P_f^{(12)} = \left( \int_{\Pi 1} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y = \boldsymbol{a}) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} - \int_{\Pi 2} Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y = \boldsymbol{a}) d\boldsymbol{f}^{(1)} d\boldsymbol{f}^{(2)} \right) + P_f^{(1)}$$

Since $P_h^{(12)}= P_h^{(1)}$ and $P_f^{(12)}= P_f^{(1)}$, we have

$$\begin{cases} \int_{\Pi 1} Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) df^{(1)} df^{(2)} = \int_{\Pi 2} Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) df^{(1)} df^{(2)} \\ \int_{\Pi 1} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right) df^{(1)} df^{(2)} = \int_{\Pi 2} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right) df^{(1)} df^{(2)} \end{cases} \quad (A.1)$$

- If $\Omega^{(12)}=\Omega^{(1)+(2)}$, then $\Psi=\Omega^{(12)}=\Omega^{(1)+(2)}$ and $\Pi 1=\Pi 2=$NULL.

$$\begin{cases} \int_{\Pi 1} Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) df^{(1)} df^{(2)} = \int_{\Pi 2} Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) df^{(1)} df^{(2)} = 0 \\ \int_{\Pi 1} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right) df^{(1)} df^{(2)} = \int_{\Pi 2} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right) df^{(1)} df^{(2)} = 0 \end{cases}$$

- If $\Omega^{(12)}\neq\Omega^{(1)+(2)}$, then $\Pi 1\neq$NULL or $\Pi 2\neq$NULL.

1) If $\Pi 1\neq$NULL and $\Pi 2\neq$NULL.

Since $\Pi 1\subseteq\Omega^{(12)}$ and $\Pi 2\not\subset\Omega^{(12)}$, $L(f^{(1)}, f^{(2)} \parallel y\neq a, y=a)>\beta^{(12)}$ holds in $\Pi 1$ but does not hold in $\Pi 2$, *i.e.*

$$\begin{cases} Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) > \beta^{(12)} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right); \left(f^{(1)}, f^{(2)}\right) \in \Pi 1 \\ Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) \leq \beta^{(12)} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right); \left(f^{(1)}, f^{(2)}\right) \in \Pi 2 \end{cases}$$

accordingly,

$$\begin{cases} \int_{\Pi} Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) df^{(1)} df^{(2)} > \beta^{(12)} \int_{\Pi} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right) df^{(1)} df^{(2)} \\ \int_{\Pi 2} Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) df^{(1)} df^{(2)} \leq \beta^{(12)} \int_{\Pi 2} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right) df^{(1)} df^{(2)} = \beta^{(12)} \int_{\Pi} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right) df^{(1)} df^{(2)} \end{cases}$$

*i.e.*

$$\int_{\Pi 1} Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) df^{(1)} df^{(2)} > \int_{\Pi 2} Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) df^{(1)} df^{(2)}$$

This contradicts with (A.1).

2) If $\Pi 1$=NULL and $\Pi 2 \neq$NULL,

$$\int_{\Pi 2} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right) df^{(1)} df^{(2)} = \int_{\Pi 1} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right) df^{(1)} df^{(2)} = 0$$

Hence, $Pr(f^{(1)}, f^{(2)} \mid y=a)=0$ in $\Pi 2$.

Since $\Pi 2 \not\subset \Omega^{(12)}$, $L(f^{(1)}, f^{(2)} \parallel y \neq a, y=a) > \beta^{(12)}$ in $\Pi 2$ does not hold, *i.e.*

$$Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) \leq \beta^{(12)} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right) = 0; \left(f^{(1)}, f^{(2)}\right) \in \Pi 2$$

Hence, $Pr(f^{(1)}, f^{(2)} \mid y \neq a)=0$ in $\Pi 2$.

In conclusion, $Pr(f^{(1)}, f^{(2)} \mid y=a)=0$ and $Pr(f^{(1)}, f^{(2)} \mid y \neq a)=0$ in $\Pi 2$. This contradicts with the assumption that at least one of the likelihood distributions is not zero at all the points in $\Omega^{(1)+(2)}$.

3) If $\Pi 1 \neq$NULL and $\Pi 2$=NULL,

$$\int_{\Pi 1} Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) df^{(1)} df^{(2)} = \int_{\Pi 2} Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) df^{(1)} df^{(2)} = 0$$

Hence, $Pr(f^{(1)}, f^{(2)} \mid y \neq a)=0$ in $\Pi 1$.

Since $\Pi 1 \subseteq \Omega^{(12)}$, $L(f^{(1)}, f^{(2)} \parallel y \neq a, y=a) > \beta^{(12)}$ holds in $\Pi 1$. This is

$$\beta^{(12)} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right) < Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) = 0; \left(f^{(1)}, f^{(2)}\right) \in \Pi 1$$

This is impossible since $\beta^{(12)} \geq 0$ and $Pr(f^{(1)}, f^{(2)} \mid y=a) \geq 0$.

In conclusion, if points with both the likelihood distributions equal to zero are not taken into consideration, $ROC(f^{(1)} \parallel y \neq a, y=a) = ROC(f^{(1)}, f^{(2)} \parallel y \neq a, y=a)$ unless $\Omega^{(1)} = \Omega^{(1)+(2)}$.

Now, we assume that $L(f^{(1)} \parallel y \neq a, y=a) = L(f^{(1)}, f^{(2)} \parallel y \neq a, y=a)$ does not always hold, *i.e.* there exist $f_0^{(2)}$ and $f_1^{(2)}$, $f_0^{(2)} \neq f_1^{(2)}$, and

$$\frac{Pr\left(f^{(1)}, f_0^{(2)} \mid y \neq a\right)}{Pr\left(f^{(1)}, f_0^{(2)} \mid y = a\right)} < \frac{Pr\left(f^{(1)}, f_1^{(2)} \mid y \neq a\right)}{Pr\left(f^{(1)}, f_1^{(2)} \mid y = a\right)}$$

Therefore, we can find a threshold $\beta$, and

$$\frac{Pr\left(f^{(1)}, f_0^{(2)} \mid y \neq a\right)}{Pr\left(f^{(1)}, f_0^{(2)} \mid y = a\right)} < \beta < \frac{Pr\left(f^{(1)}, f_1^{(2)} \mid y \neq a\right)}{Pr\left(f^{(1)}, f_1^{(2)} \mid y = a\right)}$$

This threshold separates points $(f^{(1)}, f_0^{(2)})$ and $(f^{(1)}, f_1^{(2)})$ onto different vertices on the edge of the convex hull $ROC(f^{(1)}, f^{(2)} \parallel y \neq a, y=a)$. This contradicts with $\Omega^{(12)}=\Omega^{(1)+(2)}$. So $ROC(f^{(1)} \parallel y \neq a, y=a)=ROC(f^{(1)}, f^{(2)} \parallel y \neq a, y=a)$ unless $L(f^{(1)} \parallel y \neq a, y=a)=L(f^{(1)}, f^{(2)} \parallel y \neq a, y=a)$.□

## *Proof for Theorem* 6.2:

Given two likelihood distributions $Pr(f^{(1)}, f^{(2)} \mid y \neq a)$ and $Pr(f^{(1)}, f^{(2)} \mid y=a)$, $ROC(f^{(1)}, f^{(2)} \parallel y \neq a, y=a)$ generated according to the Neyman-Pearson procedure is not under any achievable *ROC* curve at every point in the *ROC* space. So, it remains to prove that the curve $ROC(f^{(1)} \parallel y \neq a, y=a)$ is achievable given $Pr(f^{(1)}, f^{(2)} \mid y \neq a)$ and $Pr(f^{(1)}, f^{(2)} \mid y=a)$.

Since $ROC(f^{(1)} \parallel y \neq a, y=a)$ is also calculated according to the Neyman-Pearson procedure given two likelihood distributions $Pr(f^{(1)} \mid y \neq a)$ and $Pr(f^{(1)} \mid y=a)$, we have

$$\begin{cases} P_h = \int\limits_{L\left(f^{(1)}\ y \neq a, y=a\right) > \beta} Pr\left(f^{(1)} \mid y \neq a\right) df^{(1)} = \int\limits_{L\left(f^{(1)}\ y \neq a, y=a\right) > \beta} \left( \int\limits_{-\infty}^{+\infty} Pr\left(f^{(1)}, f^{(2)} \mid y \neq a\right) df^{(2)} \right) df^{(1)} \\ P_f = \int\limits_{L\left(f^{(1)}\ y \neq a, y=a\right) > \beta} Pr\left(f^{(1)} \mid y = a\right) df^{(1)} = \int\limits_{L\left(f^{(1)}\ y \neq a, y=a\right) > \beta} \left( \int\limits_{-\infty}^{+\infty} Pr\left(f^{(1)}, f^{(2)} \mid y = a\right) df^{(2)} \right) df^{(1)} \end{cases}$$

where

$$L\left(\boldsymbol{f}^{(1)} \,||\, y \neq \boldsymbol{a}, y = \boldsymbol{a}\right) = \frac{Pr\left(\boldsymbol{f}^{(1)} \mid y \neq \boldsymbol{a}\right)}{Pr\left(\boldsymbol{f}^{(1)} \mid y = \boldsymbol{a}\right)} = \frac{\int_{-\infty}^{+\infty} Pr\left(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq \boldsymbol{a}\right) d\boldsymbol{f}^{(2)}}{\int_{-\infty}^{+\infty} Pr\left(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y = \boldsymbol{a}\right) d\boldsymbol{f}^{(2)}}$$

In the above equations, $(-\infty, +\infty)$ denotes no constraint for variables. So, the curve $ROC(\boldsymbol{f}^{(1)} \,||\, y \neq \boldsymbol{a}, y = \boldsymbol{a})$ is also achievable given $Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y \neq \boldsymbol{a})$ and $Pr(\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)} \mid y = \boldsymbol{a})$. □

# References


[1] J. M. Wolfe, "Visual search," in *Attention*, H. Pashler, Ed., London, UK: University College London Press, 1996.

[2] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97-136, 1980.

[3] P. Grenier, I. Mourey-Gerosa, M. W. Brauner, A. N. Leung, S. Lenoir, M. P. Cordeau, and B. Mazoyer, "Abnormalities of the airways and lung parenchyma in asthmatics: CT observations in 50 patients and inter- and intraobserver variability," *European Radiology*, vol. 6, pp. 199-206, 1996.

[4] G. Z. Yang, L. Dempere-Marco, X. P. Hu, and A. Rowe, "Visual search: psychophysical models and practical applications," *Image and Vision Computing*, vol. 20, pp. 291-305, 2002.

[5] L. Dempere-Marco, X. P. Hu, S. L. S. MacDonald, S. M. Ellis, D. M. Hansell, and G. Z. Yang, "The use of visual search for knowledge gathering in image decision support," *IEEE Transactions on Medical Imaging*, vol. 21, no. 7, pp. 741-754, 2002.

[6] D. Noton and L. W. Stark, "Scanpaths in Eye Movements during Pattern Perception," *Science*, vol. 17, no. 1, pp. 308-311, 1971.

[7] A. L. Yarbus, *Eye Movements and Vision*. New York: Plenum Press, 1967.

[8] C. F. Nodine, H. L. Kundel, S. C. Lauver, and L. C. Toto, "Nature of expertise in searching mammograms for breast masses," *Academic Radiology*, vol. 3, no. 12, pp. 1000-1006, 1996.

[9] C. F. Nodine, H. L. Kundel, C. Mello-Thoms, S. P. Weinstein, S. G. Orel, D. C. Sullivan, and E. F. Conant, "How experience and training influence mammography expertise," *Academic Radiology*, vol. 6, pp. 575-85, 1999.

[10] G. Sperling and B. A. Dosher, "Strategy and optimization in human information processing," in *Handbook of Human Perception and performance*, vol. 1, K. Boff, L. Kaufman, and J. Thomas, Eds., New York: Wiley, 1986.

[11] C. M. Privitera and L. W. Stark, "Algorithm for Defining Visual Regions-of-Interest: Comparison with Eye Fixations," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 9, pp. 970-982, 2000.

[12] H. D. Tagare, K. Toyama, and J. G. Wang, "A maximum-likelihood strategy for directing attention during visual search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 5, pp. 490-500, 2001.

[13] W. N. Klarquist and A. C. Bovik, "FOVEA: A foveated vergent active stereo vision system for dynamic three-dimensional scene recovery," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 5, pp. 755-770, 1998.

[14] J. S. DeBonet and P. Viola, "Structure driven image database retrieval," *Advances in neural information processing*, vol. 10, 1997.

[15] G. Z. Yang, D. M. Hansell, and D. F. Gillies, "ViTAL-Visual Search for Active Learning," EPSRC (GR/N08810) Proposal, UK, 1999.

[16] D. H. Hubel, *Eye, Brain, and Vision*. New York: W.H. Freeman, 1995.

[17] N. V. Graham, *Visual Pattern Analyzers*. New York: Oxford University Press, 1989.

[18] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *Journal of Physiology*, vol. 160, pp. 106-154, 1962.

[19] D. H. Hubel and T. N. Wiesel, "Functional architecture of macaque monkey visual cortex," *Proceedings of the Royal Society of London*, B, vol. 198, pp. 1-59, 1977.

[20] R. N. Carlson, *Physiology of Behaviour*, 6th ed. Boston: Allyn And Bacon, 1998.
[21] A. D. Milner and M. A. Goodale, *The Visual Brain in Action*. Oxford: Oxford University Press, 1995.
[22] M. J. Webster and L. G. Ungerleider, "Neuroanatomy of Visual Attention," in *The Attentive Brain*, R. Parasuraman, Ed., Cambridge, Massachusetts: MIT Press, 1998, pp. 19-34.
[23] L. G. Ungerleider and M. Mishkin, "Two cortical visual systems," in *Analysis of visual behavior*, D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, Eds., Cambridge, Massachusetts: MIT Press, 1982, pp. 549-586.
[24] J. D. Cohen, K. Dunbar, and J. L. McClelland, "On the control of automatic processes: A parallel distributed processing model of the Stroop effect," *Psychological Review*, vol. 97, no. 3, pp. 332-361, 1990.
[25] D. Servan-Schreiber, H. Printz, and J. D. Cohen, "A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior," *Science*, vol. 249, pp. 892-895, 1990.
[26] J. D. Cohen, T. S. Braver, and R. C. O'Reilly, "A computational approach to prefrontal cortex, cognitive control, and schizophrenia: Recent developments and current challenges," *Philosophical Transactions of the Royal Society of London*, B, Biological Sciences, vol. 351, pp. 1515-1527, 1996.
[27] E. K. Miller and J. D. Cohen, "An integrative theory of prefrontal cortex function," *Annual Review of Neuroscience*, vol. 24, pp. 167-202, 2001.
[28] E. K. Miller, "The prefrontal cortex and cognitive control," *Nature Reviews Neuroscience*, vol. 1, pp. 59-65, 2000.
[29] H. Davson, *Physiology of the Eye*, 4th ed. New York, NY: Academic Press, 1980.
[30] S. P. Liversedge and J. M. Findlay, "Saccadic eye movements and cognition," *Trends in Cognitive Science*, vol. 4, pp. 6-14, 2000.
[31] D. P. Mounoz and R. H. Wurtz, "Fixation cells in monkey superior colliculus: I. II. characteristics of cell discharge," *Journal of Neurophysiology*, vol. 70, pp. 559-589, 1993.
[32] R. H. S. Carpenter, *Movements of the eyes*. London: Pion Limited, 1977.
[33] A. C. Guyton and J. E. Hall, *Textbook of Medical Physiology*, 9th ed. Philadelphia: W.B. Saunders, 1996.
[34] G. Smith and D. A. Atchison, *The Eye and Visual Optical Instrument*. New York: Cambridge University Press, 1997.
[35] G. Westheimer, "Visual Acuity," in *Adler's Physiology of the eye, Clinical Application, Chapter 17*, R. A. Moses and W. M. Hart, Eds., 8th ed. St. Louis: The C.V. Mosby Company, 1987.
[36] I. L. Bailey and J. E. Lovie, "New design principles for visual acuity letter charts," *American Journal of Optometry and Physiological Optics*, vol. 53, pp. 740-745, 1976.
[37] R. N. Haber and M. Hershenson, *The Psychology of Visual Perception*. New York, NY: Holt, Rinehart, and Winston, Inc., 1973.
[38] M. D'Zmura, "Bergmann on visual resolution," *Perception*, vol. 25, pp. 1223-1234, 1996.
[39] D. R. Williams and N. J. Coletta, "Cone spacing and the visual resolution limit," *Journal of the Optical Society of America*, A:4, pp. 1514-1523, 1987.
[40] D. G. Green, "Regional varitations in the visual acuity for interference fringes on the retina," *The Journal of Physiology*, vol. 207, pp. 351-356, 1970.
[41] S. J. Waugh and D. M. Levi, "Spatial alignment across gaps: contributions of orientation and spatial scale," *Journal of the Optical Society of America*, vol. 12, pp. 2305-2317, 1995.
[42] M. H. Hennig and F. Wörgötter, "Eye micro-movements improve stimulus detection beyond the Nyquist limit in the peripheral retina," in *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press, 2004.

[43] H. L. Averill and F. W. Weymouth, "Visual perception and the retinal mosaic. II. The influence of eye-movements on the displacement threshold," *Journal of Comparative Psychology*, vol. 5, pp. 147-176, 1925.

[44] M. I. Posner, C. R. R. Snyder, and B. J. Davidson, "Attention and the detection of signals," *Journal of Experimental Psychology*, General, vol. 109, pp. 160-174, 1980.

[45] M. I. Posner, "Orienting of attention," *Quarterly Journal of Experimental Psychology*, vol. 32, pp. 3-25, 1980.

[46] S. Kastner, M. Pinsk, P. D. Weerd, R. Desimone, and L. Ungerleider, "Increased activity in human visual cortex during directed attention in the absence of visual stimulation," *Neuron*, vol. 22, pp. 751-761, 1999.

[47] R. Klein, "Inhibitory tagging system facilitates visual search," *Nature*, vol. 334, pp. 430-431, 1988.

[48] K. O'Craven and N. Kanwisher, "Mental imagery of faces and places activates corresponding stimulus-specific brain regions," *Journal of Cognitive Neuroscience*, vol. 12, pp. 1013-1023, 2000.

[49] K. O'Craven and N. Kanwisher, "Extrastriate activity during visual imagery is content specific," *Cognitive Neuroscience Society Abstracts*, pp. 91, 1999.

[50] K. O'Craven and N. Kanwisher, "Visual imagery of moving stimuli activates area MT/MST," the 27th Annual Meeting of the Society for Neuroscience, vol. 23(1-2), New Orleans, LA., 1997, pp. 299.

[51] R. L. Buchner and M. E. Wheeler, "The cognitive neuroscience of remembering," *Nature Review Neuroscience*, vol. 2, pp. 624-634, 2001.

[52] S. Ogawa, T. M. Lee, A. S. Nayak, and P. Glynn, "Oxygenation-Sensitive Contrast in Magnetic Resonance Image of Rodent Brain at High Magnetic Fields," *Magnetic Resonance Medicine*, vol. 14, pp. 68-78, 1990.

[53] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proceedings of the National Academy of Sciences, USA*, vol. 87, pp. 9868-9872, 1990.

[54] S. Ogawa, R. S. Menon, D. W. Tank, S. G. Kim, H. Merkle, J. M. Ellermann, and K. Ugurbil, "Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging," *Biophysical Journal*, vol. 64, no. 3, pp. 803-812, 1993.

[55] M. S. Cohen and S. Y. Bookheimer, "Localization of brain function using magnetic resonance imaging," *Techniques in Neuroscience*, vol. 17, no. 7, pp. 268-277, 1994.

[56] L. R. Young and D. Sheena, "Methods and Designs: Survey of eye movement recording methods," *Behaviour Research Methods and Instrumentation*, vol. 7, pp. 397-429, 1975.

[57] H. D. Crane, "The Purkinje Image Eyetracker, Image Stabilization, and Related Forms of Stimulus Manipulation," in *Visual Science and Engineering: Models and Applications*, D. H. Kelly, Ed., New York, NY: Marcel Dekker, Inc., 1994, pp. 13-89.

[58] S. J. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, pp. 520-522, 1996.

[59] R. Desimone, E. K. Miller, L. Chelazzi, and A. Lueschow, "Multiple memory systems in visual cortex," in *The Cognitive Neurosciences*, M. S. Gazzaniga, Ed., Cambridge, Massachusetts: MIT Press, 1995, pp. 475-486.

[60] R. VanRullen and S. J. Thorpe, "The time course of visual processing: from early perception to decision making," *Journal of Cognitive Neuroscience*, vol. 13, no. 4, pp. 454-461, 2001.

[61] B. C. Motter, "Focal attentional produces spatially selective processing in visual areas V1, V2 and V4 in the presence of competing stimuli," *Journal of Neurophysiology*, vol. 70, pp. 909-919, 1993.

[62] S. J. Luck, L. Chelazzi, S. A. Hillyard, and R. Desimone, "Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex," *Journal of Neurophysiology*, vol. 77, pp. 24-42, 1997.

[63] S. Treue and J. C. Martinez-Trujillo, "Feature-based attention influences motion processing gain in macaque visual cortex," *Nature*, vol. 399, pp. 575-579, 1999.

[64] J. Moran and R. Desimone, "Selective attention gates visual processing in the extrastriate cortex," *Science*, vol. 229, pp. 782-784, 1985.

[65] S. Treue and J. H. R. Maunsell, "Attentional modulation of visual motion processing in cortical areas MT and MST," *Nature*, vol. 382, pp. 539-541, 1996.

[66] A. T. Smith, K. D. Singh, and M. W. Greenlee, "Attentional suppression of activity in human visual cortex," *NeuroReport*, vol. 11, pp. 271-277, 2000.

[67] J. M. Wolfe, K. R. Cave, and S. Franzel, "Guided Search: An alternative to the feature integration model for visual search," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, pp. 419-433, 1989.

[68] D. E. Broadbent, *Perception and Communication*. New York: Pergamon Press, 1958.

[69] A. M. Treisman, "Contextual cues in selective listening," *Quarterly Journal of Experimental Psychology*, vol. 12, pp. 242-248, 1960.

[70] J. A. Deutsch and D. Deutsch, "Attention: some theoretical considerations," *Psychological Review*, vol. 70, pp. 80-90, 1963.

[71] S. J. Luck and M. Girelli, "Electrophysiological approaches to the study of selective attention in the human brain," in *The Attentive Brain*, R. Parasuraman, Ed., Cambridge, Massachusetts: MIT Press, 1998, pp. 71-94.

[72] M. Ito and C. D. Gilbert, "Attention modulates contextual influences in the primary visual cortex of alert monkeys," *Neuron*, vol. 22, pp. 593-604, 1999.

[73] S. P. Gandhi, D. J. Heeger, and G. M. Boynton, "Spatial attention affects brain activity in human primary visual cortex," *Proceedings of the National Academy of Sciences, USA*, vol. 96, pp. 3314-3319, 1999.

[74] A. Martinez, L. Anllo-Vento, M. I. Sereno, L. R. Frank, R. B. Buxton, D. J. Dubowitz, E. C. Wong, H. Hinrichs, H. J. Heinze, and S. A. Hillyard, "Involvement of striate and extrastriate visual cortical areas in spatial attention," *Nature Neuroscience*, vol. 2, pp. 364-369, 1999.

[75] K. M. O'Craven, P. E. Downing, and N. Kanwisher, "fMRI evidence for objects as the units of attentional selection," *Nature*, vol. 401, pp. 584-587, 1999.

[76] P. R. Roelfsema, V. A. F. Lamme, and H. Spekreijse, "Object-based attention in the primary visual cortex of the macaque monkey," *Nature*, vol. 395, pp. 376-381, 1998.

[77] N. Lavie, "Perceptual load as a necessary condition for selective attention," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21, pp. 451-468, 1995.

[78] G. Rees, C. D. Frith, and N. Lavie, "Modulation irrelevant motion perception by varying attentional load in an unrelated task," *Science*, vol. 28, pp. 1616-1619, 1997.

[79] T. C. Handy and G. R. Mangun, "Attention and spatial selection: electrophysiological evidence for modulation by perceptual load," *Perception and Psychophysics*, vol. 62, pp. 175-186, 2000.

[80] C. E. Connor, D. C. Preddie, J. L. Gallant, and D. C. Van-Essen, "Spatial attention effects in macaque area V4," *Journal of Neuroscience*, vol. 17, pp. 3201-3214, 1997.

[81] J. Duncan, "Selective attention and the organization of visual information," *Journal of Experimental Psychology*, General, vol. 113, pp. 501-17, 1984.

[82] J. M. Wolfe and C. W. Pokorny, "Inhibitory tagging in visual search: A failure to replicate," *Perception and Psychophysics*, vol. 48, pp. 357-362, 1990.

[83] R. Hubner and G. Backer, "Perceiving spatially inseparable objects: Evidence for feature-based object selection not mediated by location," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 25, pp. 1556-1567, 1999.
[84] J. E. Hoffman, "A two-stage model of visual search," *Perception and psychophysics*, vol. 25, pp. 319-327, 1979.
[85] M. J. Nichols and W. T. Newsome, "The neurobiology of cognition," *Nature*, vol. 402, no. 2, pp. C35-C38, 1999.
[86] N. Moray, "The role of attention in the detection of errors and the diagnosis of errors in man machine systems," in *Human Detection and Diagnosis of Failures*, J. Rasmussen and W. Rouse, Eds., New York: Plenum Press, 1981.
[87] N. Moray, "Monitoring behaviour and supervisory control," in *Handbook of human Perception and performance*, K. Boff, L. Kaufmann, and J. Beattie, Eds., New York: Wiley, 1986.
[88] J. E. Russo and D. R. Larry, "An Eye Fixation Analysis of Multi-alternative Choice," *Memory and Cognition*, vol. 3, pp. 267-276, 1975.
[89] E. D. Megaw and J. Richardson, "Target uncertainty and visual scanning strategies," *Human Factors*, vol. 21, pp. 303-316, 1979.
[90] R. E. Morrison, "Manipulation of stimulus onset delay in reading: evidence for parallel programming of saccades," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 10, pp. 667-682, 1984.
[91] N. Graham, "Breaking the visual stimulus into parts," *Current Directions in Psychological Science*, vol. 1, pp. 55-61, 1992.
[92] F. W. Campbell and J. G. Robson, "Application of Fourier analysis to the visibility of gratings," *Journal of Physiology*, vol. 197, pp. 551-566, 1968.
[93] N. Kanwisher, J. McDermott, and M. Chun, "The fusiform face area: a module in human extrastriate cortex specialized for the perception of faces," *Journal of Neuroscience*, vol. 17, pp. 4302-4311, 1997.
[94] R. Epstein and N. Kanwisher, "A cortical representation of the local visual environment," *Nature*, vol. 392, pp. 598-601, 1998.
[95] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual Review of Neuroscience*, vol. 18, pp. 193-222, 1995.
[96] N. Kanwisher and E. Wojciulik, "Visual Attention: Insights from Brain Imaging," *Nature Reviews Neuroscience*, vol. 1, pp. 91-100, 2000.
[97] J. B. Hopfinger, M. H. Buonocore, and G. R. Mangun, "The neural mechanisms of top-down attentional control," *Nature Neuroscience*, vol. 3, pp. 284-291, 2000.
[98] J. L. Bradshaw and J. B. Mattingley, "Clinical Neuropsychology," in *Behavioral and Brain Sciences, Chapter 6*. San Diego: Academic Press, 1995.
[99] K. M. Heilman and E. Valenstein, *Clinical Neuropsychology, Chapter 10*, 3rd ed. New York: Oxford Press, 1993.
[100] M. M. Sohlberg and C. A. Mateer, *Introduction to Cognitive Rehabilitation: Theory and Practice*. New York: Guilford Press, 1989.
[101] R. A. Rensink, J. K. O'Regan, and J. J. Clark, "To See or Not to See: The Need for Attention to Perceive Changes in Scenes," *Psychological Science*, vol. 8, pp. 368-373, 1997.
[102] R. A. Rensink, "The dynamic representation of scenes," *Visual Cognition*, vol. 7, pp. 17-42, 2000.
[103] D. J. Simons, "In sight, out of mind: When object representations fail," *Psychological Science*, vol. 7, pp. 301-305, 1996.
[104] D. J. Simons and D. T. Levin, "Change Blindness," *Trends in Cognitive Sciences*, vol. 1, pp. 261-267, 1997.
[105] D. J. Simons, "Current approaches to change blindness," *Visual Cognition*, vol. 7, pp. 1-15, 2000.
[106] T. S. Horowitz and J. M. Wolfe, "Visual search has no memory," *Nature*, vol. 394, pp. 575-577, 1998.

[107] I. D. Gilchrist and M. Harvey, "Strategic scanning as a substitute for memory in visual search," Conference Program and Abstract of the 11th European Conference on Eye Movements, Turku, Finland, 2001, pp. S10.
[108] L. Itti and C. Koch, "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience*, vol. 2, pp. 194-203, 2001.
[109] M. Riesenhuber and T. Poggio, "Computation models of object recognition in cortex: a review," MIT report, A.I. Memo No. 1695, C.B.C.L Paper No. 190, the Center for Biological and Computational Learning, the Department of Brain and Cognitive Sciences, the Massachusetts Institute of Technology, August, 2000.
[110] K. Fukushima, "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," *Biological Cybernetics*, vol. 36, pp. 193-202, 1980.
[111] R. P. N. Rao and D. H. Ballard, "Dynamic model of visual recognition predicts neural response properties in the visual cortex," *Neural Computation*, vol. 9, pp. 721-763, 1997.
[112] I. Biederman, "Recognition-by-components: A theory of human image understanding," *Psychological Review*, vol. 94, pp. 115-117, 1987.
[113] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219-227, 1985.
[114] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489-1506, 2000.
[115] D. Marr, "Early processing of visual information," *Philosophical Transactions of the Royal Society of London*, Series B, pp. 483-524, 1976.
[116] J. M. H. du-Buf, "Gabor phase in texture discrimination," *Signal Processing*, vol. 21, pp. 221-240, 1990.
[117] K. Raner and A. Pollatsek, "Eye movement and scene perception," *Canadian Journal of psychology*, vol. 46, no. 3, pp. 342-376, 1992.
[118] M. C. Mozer and M. Sitton, "Computation modeling of spatial attention," in *Attention*, H. Pashler, Ed., London: UCL press, 1996.
[119] C. L. Colby and M. E. Goldberg, "Space and attention in parietal cortex," *Annual Review of Neuroscience*, vol. 22, pp. 319-349, 1999.
[120] J. D. Schall and K. G. Thompson, "Neural selection and control of visually guided eye movements," *Annual Review of Neuroscience*, vol. 22, pp. 241-259, 1999.
[121] J. M. Findlay, "Global processing for saccadic eye movements," *Vision Research*, vol. 22, pp. 1033-1045, 1982.
[122] P. He and E. Kowler, "The role of location probability in the programming of saccades: implications for center-of-gravity tendencies," *Vision Research*, vol. 29, pp. 1165-1181, 1989.
[123] R. P. N. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard, "Eye Movements in Iconic Visual Search," *Vision Research*, vol. 42, no. 11, pp. 1447-1463, 2002.
[124] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, and C. Zetzsche, "Scene analysis with saccadic eye movements: top-down and bottom-up modeling," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 152-160, 2001.
[125] T. S. Lee and S. Yu, "An information-theoretic framework for understanding saccadic eye movements," in *Advance in Neural Information Processing Systems*, vol. 12. Cambridge, Massachusetts: MIT Press, 2000.
[126] A. Oliva and P. G. Schyns, "Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli," *Cognitive Psychology*, vol. 34, pp. 72-107, 1997.
[127] J. M. Findlay, D. Brogan, and S. M. Wenban, "The visual signal for saccadic eye movements emphasizes visual boundaries," *Perception and Psychophysics*, vol. 53, pp. 633-641, 1993.
[128] T. Lindeberg, *Scale-Space Theory in Computer Vision*. Boston: Kluwer Academic Publishers, 1994.

[129] S. S. Hacisalihzade and L. W. Stark, "Visual perception and sequences of eye movement fixations: A stochastic modeling approach," *IEEE Transactions on System, Man, and Cybernetics*, vol. 22, no. 3, pp. 474-481, 1992.
[130] D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*. New York: John Wiley and Sons, 1966.
[131] G. A. Sperling, "Unified theory of attention and signal detection," in *Varieties of attention*, R. Parasuraman and D. R. Davies, Eds., New York: Academic Press, 1984, pp. 103-181.
[132] P. Verghese, "Visual search and attention: A signal detection theory approach," *Neuron, Review*, vol. 31, pp. 523-535, 2001.
[133] J. Palmer, P. Verghese, and M. Pavel, "The psychophysics of visual search," *Vision Research*, vol. 40, pp. 1227-1268, 2000.
[134] R. A. Young, "The Gaussian derivative model for spatial vision: I. Retinal mechanisms," *Spatial Vision*, vol. 2, pp. 273-293, 1987.
[135] B. Van-Ginneken and B. M. ter-Haar-Romeny, "Applications of Locally Orderless Images," *Journal of Visual Communication and Image Representation*, vol. 11, no. 2, pp. 196-208, 2000.
[136] F. Chabat, "Computed tomography image analysis for the detection of obstructive diseases," PhD. Dissertation, National Heart and Lung Institute, Imperial College of Science, Technology and Medicine, University of London (London, U.K.), 2000.
[137] ASL, "Eye Tracking System Instruction Manual-Software. Manual Version 1.2," ASL Applied Science Laboratories, 2000.
[138] S. Martinez-Conde, S. L. Macknik, and D. H. Hubel, "The role of fixational eye movements in visual perception," *Nature Reviews Neuroscience*, vol. 5, pp. 229-240, 2004.
[139] J. T. Fulton, "Dynamics of the physiological optics of vision," in *Processes in biological vision, chapter 7.3*, 2003. http://www.4colorvision.com/.
[140] J. M. Findlay, "Frequency analysis of human involuntary eye movement," *Kybernetik*, vol. 8, pp. 207-14, 1971.
[141] J. M. Reinhardt, N. D. D'Souza, and E. A. Hoffman, "Accurate Measurement of Intrathoracic Airways," *IEEE Transactions on Medical Imaging*, vol. 16, no. 5, pp. 820-827, 1997.
[142] F. Chabat, X. P. Hu, D. M. Hansell, and G. Z. Yang, "ERS Transform for the Automated Detection of Bronchial Abnormalities on CT of the Lungs," *IEEE Transactions on Medical Imaging*, vol. 20, no. 9, pp. 942-952, 2001.
[143] G. H. Landeweerd and E. S. Gelsema, "The Use of Nuclear Texture Parameters in the Automatic Analysis of Leukocytes," *Pattern Recognition*, vol. 10, pp. 57-61, 1978.
[144] D. Chappard, A. Chennebault, M. Moreau, E. Legrand, M. Audran, and M. F. Basle, "Texture Analysis of X-Ray Radiographs Is a More Reliable Descriptor of Bone Loss Than Mineral Content in a Rat Model of Localized Disuse Induced by the Clostridium botulinum Toxin," *Bone*, vol. 28, no. 1, pp. 72-79, 2001.
[145] F. Chabat, G. Z. Yang, and D. M. Hansell, "Obstructive lung diseases: texture classification for differentiation at CT," *Radiology*, vol. 228, pp. 871-877, 2003.
[146] K. Karu, A. K. Jain, and R. M. Bolle, "Is there any texture in the image?," *Pattern recognition*, vol. 29, no. 9, pp. 1437-1446, 1996.
[147] R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of IEEE*, vol. 67, pp. 786-804, 1979.
[148] K. I. Laws, "Textured Image Segmentation," PhD thesis, Image Processing Institute, University of Southern California, 1980.
[149] A. K. Jain and K. Karu, "Learning texture discrimination masks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 195-205, 1996.

[150] L. van-Gool, P. Dewaele, and A. Oosterlinck, "Texture analysis: Anno 1983," *Computer Vision, Graphics and Image Processing*, vol. 29, pp. 336-357, 1985.
[151] M. Tuceryan and A. K. Jain, "Texture analysis," in *Handbook Pattern Recognition and Computer Vision, Chapter 2*, C. H. Chen, L. F. Pau, and P. S. P. Wang, Eds., Singapore: World Scientific, 1993, pp. 235-276.
[152] T. R. Reed and J. M. H. Buf, "A review of recent texture segmentation and feature extraction techniques," *Computer Vision Graphics and Image Processing: Image Understanding*, vol. 57, no. 3, pp. 359-372, 1993.
[153] S. W. Zucker, "Toward a model of Texture," *Computer Graphics and image processing*, vol. 5, pp. 190-202, 1976.
[154] H. Voorhees and T. Poggio, "Detecting textons and texture boundaries in natural images," Proceedings of the First International Conference on Computer Vision, London, UK, 1987, pp. 250-258.
[155] B. Julesz, "Visual Pattern Discrimination," *IRE Transactions on Information Theory*, IT-8(2), pp. 84-92, 1962.
[156] S. R. Purks and W. Richards, "Visual Texture Discrimination Using Random-Dot Patterns," *Journal of the Optical Society of America*, A, vol. 67(6), pp. 765-771, 1977.
[157] R. M. Haralick and R. Bosley, "Texture features for image classification," Proceedings of the Third ERTS Symposium, NASA SP-351, 1973, pp. 1219-1228.
[158] J. S. Weszka, C. R. Dyer, and A. Rosenfeld, "A comparative study of texture measures for terrain classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 6, pp. 269-285, 1976.
[159] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer Graphics Image Process*, vol. 4, pp. 172-179, June 1975.
[160] H. Kaizer, "A Quantification of Textures on Aerial Photographs," Technical Note, No. 121, AD 69484, Boston University Research Laboratories, Boston University, 1955.
[161] G. R. Cross and A. K. Jain, "Markov random field texture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, pp. 25-39, 1983.
[162] H. Derin and H. Elliott, "Modeling and segmentation of noisy and textured images using Gibbs random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, pp. 39-55, 1987.
[163] R. Chellappa and S. Chatterjee, "Classification of textures using Gaussian Markov random fields," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, pp. 959-963, August 1985.
[164] B. Kartikeyan and A. Sarkar, "An identification approach for 2-D autoregressive models in describing textures," *Computer Vision, Graphics and Image Processing, Graphical Models and Image Processing*, vol. 53, pp. 121-131, March 1991.
[165] B. B. Mandelbrot, *The Fractal Geometry of Nature*. San Francisco: W.H. Freeman, 1983.
[166] A. A. Pentland, "Fractal-based description of natural scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 11, pp. 661-764, 1984.
[167] Y. Y. Tang, H. Ma, D. Xi, X. G. Mao, and C. Y. Suen, "Modified Fractal Signature (MFS): A New Approach to Document Analysis for Automatic Knowledge Acquisition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 9, no. 5, pp. 747-762, 1997.
[168] J. Huang and D. L. Turcotte, "Fractal image analysis: application to the topography of Oregon and synthetic images," *Journal of the Optical Society of America*, A, vol. 7, pp. 1124-1130, 1990.
[169] K. R. Castleman, *Digital Image Processing*. Englewood Cliffs, NJ: Prentice Hall, 1996.

[170] R. Bajcsy, "Computer Identification of Visual Surface," *Computer Graphics and Image Processing*, vol. 2, pp. 118-130, 1973.
[171] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," *Vision Research*, vol. 20, pp. 847-856, 1980.
[172] M. Clark and A. C. Bovik, "Texture segmentation using Gabor modulation/demodulation," *Pattern Recognition Letters*, vol. 6, pp. 261-267, 1987.
[173] P. P. Ohanian and R. C. Dubes, "Performance Evaluation for Four Classes of Textural features," *Pattern Recognition*, vol. 25, pp. 819-833, 1992.
[174] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis and machine vision*. London: PWS Publishing, 1999.
[175] K. Tieu and P. Viola, "Boosting Image Retrieval," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, Hilton Head Island, 2000, pp. 228-235.
[176] J. Tenembaum, V. de-Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290(5500), pp. 2319-2323, 2000.
[177] S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290(5500), pp. 2323-2326, 2000.
[178] K. P. White, T. L. Hutson, and T. E. Hutchinson, "Modeling human eye behavior during mammographic scanning: preliminary results," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 27, no. 4, pp. 494-505, July 1997.
[179] H. L. Kundel, C. F. Nodine, and E. A. Krupinski, "Searching for lung nodules: visual dwell indicates locations of false-positive and false-negative decisions," *Investigative Radiology*, vol. 24, no. 6, pp. 472-478, 1989.
[180] H. L. Kundel, C. F. Nodine, D. Thickman, and L. Toto, "Searching for lung nodules: a comparison of human performance with random and systematic scanning models," *Investigative Radiology*, vol. 22, no. 5, pp. 417-422, 1987.
[181] M. Pomplun, W. M. Reingold, J. Shen, and D. E. Williams, "The area activation model of saccadic selectivity in visual search," Proceedings of the Twenty Second Annual Conference of the Cognitive Science Society, Mahwah, NJ, 2000, pp. 375-380.
[182] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79-86, 1951.
[183] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, 1998.
[184] J. Intriligator and P. Cavanagh, "The spatial resolution of visual attention," *Cognitive Psychology*, vol. 43, pp. 171-216, 2001.
[185] S. J. Anderson, K. T. Mullen, and R. F. Hess, "Human peripheral spatial resolution for achromatic and chromatic stimuli: limits imposed by optical and retinal factors," *Journal of Physiology, London*, vol. 442, pp. 47-64, 1991.
[186] D. Noton and L. W. Stark, "Eye Movements and Visual Perception," *Scientific American*, vol. 224, no. 6, pp. 35-43, 1971.
[187] X. P. Hu, L. Dempere-Marco, and G. Z. Yang, "Hot Spot Detection Based on Feature Space Representation of Visual Search," *IEEE Transactions on Medical Imaging*, vol. 22, no. 9, pp. 1152-1162, 2003.
[188] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245-271, 1997.
[189] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273-324, 1997.
[190] W. W. Cooley and P. R. Lohnes, *Multivariate Data Analysis*. New York: John Wiley and Sons, Inc., 1971.

[191] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," Proceedings of the 11th International Conference on Machine Learning, New Brunswick, NJ, 1994, pp. 121-129.

[192] D. Koller and M. Sahami, "Towards optimal feature selection," Proceedings of 13th International Conference on Machine Learning, Bari, Italy, 1996, pp. 284-292.

[193] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," Proceedings of the 9th National Conference on Artificial Intelligence, Cambridge, Massachusetts: MIT Press, 1992, pp. 129-134.

[194] H. Almuallim and T. G. Dietterich, "Learning with many features," Proceedings of the 9th National Conference on Artificial Intelligence, Cambridge, Massachusetts: MIT Press, 1992, pp. 547-552.

[195] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131-156, August, 1997.

[196] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI), San Mateo, CA, Morgan Kaufmann, 1995, pp. 1137-1143.

[197] J. P. Egan, *Signal detection theory and ROC analysis*. New York: Academic Press, 1975.

[198] H. L. Van-Trees, *Detection Estimation and Modulation Theory*, vol. 1-3. New York: Wiley and Sons, 1971.

[199] M. J. J. Scott, M. Niranjan, and R. W. Prager, "Parcel: feature subset selection in variable cost domains," Technical Report, CUED/F-INFENG/TR. 323, Department of Engineering, University of Cambridge, England, May, 1998.

[200] A. Srinivasan, "Note on the location of optimal classifiers in n-dimensional ROC space," Technical Report, PRG-TR-2-99, Oxford University Computing Laboratory, Oxford, England, 1999.

[201] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, pp. 80-83, 1945.

[202] D. J. Hand, *Construction and assessment of classification rules*. Chichester, England: John Wiley and Sons, 1997.

[203] J. Pearl, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann, 1988.

[204] F. Coetzee, S. Lawrence, and C. L. Giles, "Bayesian classification and feature selection from finite data sets," Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence (UAI-2000), Stanford, CA, July 2000, pp. 89-97.

[205] M. Singh and G. M. Provan, "Efficient learning of selective Bayesian network classifiers," Proceedings of 13th International Conference on Machine Learning, Bari, Italy, 1996, pp. 453-461.

[206] C. L. Blake and C. J. Merz, UCI Repository of machine learning databases, 1998, http://www.ics.uci.edu/~mlearn/MLRepository.html.

[207] P. Langley, W. Iba, and K. Thompson, "An analysis of Bayesian classifiers," Proceedings of the 10th National Conference on Artificial Intelligence, San Jose, CA, 1992, pp. 223-228.

[208] J. R. Quinlan, *C4.5: Programs for machine learning*. Los Altos, CA: Morgan Kaufmann, 1993.

[209] R. Kohavi and D. Sommerfield, "Data mining using MLC++: a machine learning library in C++," *IEEE Computer Society Press*, pp. 234-45, 1996.

[210] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under the receiver operating characteristic (ROC) curve," *Diagnostic Radiology*, vol. 143, pp. 29-36, 1982.

[211] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998.

[212] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696-710, 1997.
[213] A. Torralba, "Modeling global scene factors in attention," *Optical Society of America*, vol. 20, no. 7, pp. 1407-1418, July 2003.
[214] J. Duncan and G. W. Humphreys, "Visual search and stimulus similarity," *Psychological Review*, vol. 96, no. 3, pp. 433-458, 1989.
[215] M. E. Poisson and F. Wilkinson, "Distractor ratio and grouping processes in visual conjunction search," *Perception*, vol. 21, pp. 21-38, 1992.
[216] H. C. Arsenio, A. Oliva, and J. M. Wolfe, "'Exorcising 'ghosts' in repeated visual search," *Journal of Vision*, vol. 2, no. 7, pp. 733a, 2002.
[217] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *Society for Industrial and Applied Mathematics Review*, vol. 26, no. 2, pp. 195-239, 1984.
[218] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, Series B39, pp. 1-38, 1977.
[219] S. Akaho, "The EM algorithm for multiple object recognition," Proceedings of IEEE International Conference on Neural Networks (ICNN'95), Perth, Western Australia, 1995, pp. 2426-2431.
[220] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159-179, 1985.
[221] M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*, 2nd ed. New York: John Wiley and Sons, 1999.
[222] NASA's Earth Observatory. http://earthobservatory.nasa.gov/.
[223] C. F. Nodine and H. L. Kundel, "The cognitive side of visual search in radiology," in *Eye Movements: From Physiology to Cognition*, J. K. O'Regan and A. Levy-Schoen, Eds., New York: Elsevier Science, 1987, pp. 573-582.
[224] L. M. Lifshitz and S. M. Pizer, "Multiresolution hierarchical approach to image segmentation bsed on intensity extrema," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 6, pp. 529-540, 1990.
[225] M. F. Land and P. Mcleod, "From eye movements to actions: how batsmen hit the ball," *Nature Neuroscience*, vol. 3, pp. 1340-1345, 2000.
[226] L. Bretzner and T. Lindeberg, "Feature tracking with automatic selection of spatial scales," *Computer Vision and Image Understanding*, vol. 71, no. 3, pp. 385-392, 1998.
[227] G. Adiv, "Determining 3-D motion and structure from optical flow generated by several moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 7, pp. 384-401, July 1985.
[228] J. W. Roach and J. K. Agganval, "Determining the movement of objects from a sequence of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, pp. 554-562, November 1980.
[229] D. Regan and K. I. Beverly, "Binocular and monocular stimuli for motion in depth: changing-disparity and changing-size feed the same motion-in-depth stage," *Vision Research*, vol. 19, pp. 1331-1342, 1979.
[230] H. Fuchs, M. A. Livingston, R. Raskar, D. Colucci, K. Keller, A. State, J. R. Crawford, P. Rademacher, S. H. Drake, and A. A. Meyer, "Augmented reality visualization for laparoscopic surgery," Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention, Cambridge, MA, October 1998, pp. 934-943.