Digital System for Bio-inspired Visual Attention Processing Fast and Efficient Information Theoretic Modelling of Saliency

Anh Cat Le Ngo, BEng.

Thesis submitted to The University of Nottingham
for the degree of Doctor of Philosophy

July, 2015

Anh Cat Le Ngo

July, 2015

Dedicated to my family and my love

Learn from yesterday, live for today, hope for tomorrow.

The important thing is not to stop questioning

R.I.P Albert Einstein

Abstract

Visual attention is a biological mechanism of human vision systems to cope with rich and fast-changing visual information in surrounding environments. Visual saliency is a strategy, which recommends attentive spots to be visited in descending orders of interest or information amounts. This thesis aims to utilize information theory in computational saliency models, assumed that more attention is drawn toward more informative locations. As visual media, i.e. images and videos, are high-dimensional data, information estimation is often computationally infeasible due to enormous requirement of computation and data samples. This thesis proposes and analyses three different practical and innovative information-based saliency models. The first model, called entropy-based saliency method (ENT), measures salient information with centre-surrounding operation by conditional entropy (ENT-CON) or Kullback-Leibler divergence (ENT-KLD). However, ENT only estimates information from local features of fixed-size windows, it does not utilize multi-scale and global information of visual media, which are proven to be important in biological visual attention. To utilise multi-scale information, Wavelet-based Scale-Saliency (WSS), the second model, estimates information from power distribution of data across wavelet sub-bands basis descriptors in multiple dyadic scales. Though WSS has benefited from local features at multiple scales, it has not integrated information of global context or statistical characteristics of natural images. Multiscale Discriminant Saliency (MDIS), the third model, adopts Wavelet Hidden Markov Tree (WHMT) to unify both multiple-scale and global information for a comprehensive saliency method. All three models, ENT, WSS and MDIS are evaluated and compared against well-known saliency methods such as PSS, AIM, DIS, etc quantitatively by standard numerical tools (Normalized Scale Saliency (NSS), Linear Correlation Coefficient (LCC), Area Under Curver (AUC)) on N.Bruce's, Kootstra's and Judd's databases with human eye-tracking ground-truth as well as qualitatively by visual examination of individual cases. Performances and comprehensiveness of three models are reflected through numerical results of an experiment on Bruce's database, Table 1. As the latter model is designed in more comprehensive and

computationally complex manner than the previous, all three quantitative evaluations (LCC,NSS,AUC) generally and computational time increase in that order.

	ENT	WSS	MDIS
LCC	0.02263	-0.01731	0.02382
NSS	-0.17533	0.31782	0.48019
AUC	0.78167	0.70292	0.88335
TIME(s/frame)	0.87040	1.26889	2.32734

Table 1: ENT,WSS,MDIS's quantitative results on N.Bruce's database

List of Publications

[1] A. C. L. Ngo, K. L.-M. Ang, J. K.-P. Seng, and G. Qiu, "Multi-scale discriminant saliency with wavelet-based hidden markov tree modelling," *Computers & Electrical Engineering*, vol. 40, no. 4, pp. 1376 – 1389, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0045790614000275

[2] A. C. Le Ngo, K. A. Li-Minn, G. Qiu, and J. S. Kah-Phooi, "Multi-scale visual attention & saliency modelling with decision theory," in *Proceeding of International Image Processing Conference (ICIP 2013)*. Melbourne, Australia: IEEEXplore, Sep. 2013, accepted.

[3] A. C. Le Ngo, L. Ang, K. P. Seng, and G. Qiu, "Information-based scale saliency methods with wavelet sub-band energy density descriptors," in *Lecture Series of Computer Science*, ser. Lecture Notes in Computer Science, vol. 7803. Kuala Lumpur, Malaysia: Springer, 2013. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-36543-0_38

[4] A. C. Le Ngo, G. Qiu, G. Underwood, L. M. Ang, and K. P. Seng, "Visual saliency based on fast nonparametric multidimensional entropy estimation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, p. 1305–1308. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6288129

[5] A. C. Le Ngo, L. Ang, K. P. Seng, and G. Qiu, "Improvement and evaluation of visual saliency based on information theory," in *Computer Symposium (ICS), 2010 International*, Dec. 2010, pp. 500 –505. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5685461

[6] A. C. Le Ngo, L. M. Ang, K. P. Seng, and G. Qiu, "Colour-based bottom-up saliency for traffic sign detection," in *Computer Applications and Industrial Electronics (ICCAIE), 2010 International Conference on*, 2010, p. 427–431. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5735117

[7] A. C. Le Ngo, L. Ang, K. P. Seng, and G. Qiu, "Lane-mark extraction by frequency-based saliency visual attention," in *2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, vol. 9, Jul. 2010, pp. 55 –59. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5564495

[8] K. H. Lim, A. C. Le Ngo, K. P. Seng, and L. Ang, "UNMC-VIER AutoVision database," in *2010 International Conference on Computer Applications and Industrial Electronics (ICCAIE)*, Dec. 2010, pp. 650 –654. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5735015

[9] K. H. Lim, K. P. Seng, A. C. Le Ngo, and L. Ang, "Real-Time implementation of Vision-Based lane detection and tracking," in *International Conference on Intelligent Human-Machine Systems and Cybernetics, 2009. IHMSC '09*, vol. 2, Aug. 2009, pp. 364 –367. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5335965

Acknowledgements

Although being a PhD student is not the toughest job in the world, it is certainly not easy to spend several years in labs and face research work day after day. Luckily, I have never walked alone by myself on this long and difficult path since there are several very-important-people (VIPs) accompanying me through hard-ship of my PhD life. I would like to take this chance and send my sincere appreciations to my supervisors, lab-mates, colleagues, family, friends, students for your tireless support throughout those years. In addition, I would like to mention few significantly important people who have patiently followed every step of my research career so far.

I wish to express my sincere gratitude to my main supervisors, Dr. Kenneth Li-Minn Ang, Dr. Guoping Qiu and Dr. Jasmine Seng Kah-Phooi for their support and guidance during my research and completion of this dissertation. For allowing me the freedom to explore my interests with understanding, continual encouragement and intellectual support even though I have not always performed as well as expected. For providing a constructive learning environment, benefiting my intellectual development of knowledge and research skills.

To my co-supervisors Dr. Lim Wee Gin, Mr. Anandan Shanmugam for your vital support and advices during my entire PhD time.

To Prof Geoff Underwood for sharing his insights of visual attention phenomena and invaluable eye-tracking, video data of daily driving which allows through assessment of my modelling works.

To my parents and sister, for their strong, continued and firm support to realize my goal. For their love, understanding, effort and advice without which this work would not have been finished.

To my love, Phung Hai Ngoc, for her patience, caring, compassion, and love. For enriching my life with her kindness, humours and thoughtfulness and being my constant source of motivation and happiness.

To my lab-mates Dr. King Hann, Dr. Wendy, Dr. Wei Chen, Lee Seng, Milor, Sue Inn,

Dr. Wai Chong, Dr. Adam, Dr. Jonathan, Eugene, Dr. Christopher in Visual Information Engineering Group (VIER), Fu Hao, Bozhi in Visual Information Processing LAB (VIPLAB) for providing necessary distractions from works, balancing my PhD experience with extra activities and enriching my dissertation with your profound ideas and comments.

To all administrative staffs from School of Engineering, UNMC and School of Computer Sciences, UNUK for your hard-work to create world-class research environments and conditions for me and several other students.

Finally, I wish to acknowledge the financial support from MIDAS scholarship and the University of Nottingham, Malaysia Campus which allow me conduct my research in both Malaysia and United Kingdom.

Contents

List of Figures 4

List of Tables 6

List of Abbreviations 7

List of Mathematical Symbols 11

1 General Introduction 13

1.1 MOTIVATION . . . 13

1.2 SCOPE . . . 14

1.3 OBJECTIVE . . . 15

1.4 CONTRIBUTION . . . 15

1.5 ORGANIZATION . . . 16

1.6 SUMMARY . . . 17

2 Literature Review 19

2.1 INTRODUCTION . . . 19

2.2 SALIENCY CONCEPT . . . 22

2.3 SALIENCY MODELLING . . . 23

2.3.1 Visual Stimuli & Eye Movement . . . 24

2.3.2 Experimental Setup & Data Evaluation . . . 26

2.4 COMPUTATIONAL MODELS . . . 30

2.4.1 Feature-based Saliency Method . . . 32

2.4.2 Frequency-based Saliency Method . . . 36

2.4.3 Information-based Saliency Methods . . . 39
2.5 SALIENCY APPLICATIONS . . . 61
2.5.1 Feature Integration Theory for Colour-Based Traffic Sign Detection 62
2.5.2 Frequency-based Saliency Methods for Lane-mark Detection . . . 68
2.6 SUMMARY . . . 71

3 Fast non-parametric entropy-based saliency (ENT) 76
3.1 INTRODUCTION . . . 76
3.2 FAST ENTROPY-BASED SALIENCY CONCEPT . . . 77
3.3 FAST ENTROPY-BASED SALIENCY (ENT) MODEL . . . 79
3.4 ENTROPY-BASED SALIENCY FOR VIDEO . . . 83
3.5 ENTROPY-BASED SALIENCY & DATA PREPROCESSING . . . 85
3.5.1 Data Preprocessing - Dimension Reduction & Decorrelation . . . 85
3.5.2 Discrete Cosine Transform - Temporal Data Decorrelation . . . 88
3.5.3 Data Preprocessing - Image De-nosing . . . 90
3.6 DISCUSSION & RESULTS . . . 92
3.6.1 Experiments on Image Databases . . . 94
3.6.2 Experiments on Video Databases . . . 98
3.7 CONCLUSION . . . 109

4 Computational Model for Wavelet-based Scale Saliency (WSS) 111
4.1 INTRODUCTION . . . 111
4.2 SCALE SALIENCY CONCEPT . . . 113
4.3 WAVLET-BASED SCALE SALIENCY MODEL . . . 115
4.3.1 Saliency in Scale Space and Spectral Space . . . 115
4.3.2 Energy Distribution in Scale Space and Spectral Space . . . 117
4.3.3 Wavelet Transform . . . 119
4.3.4 Wavelet Coefficients Correlation . . . 130
4.3.5 Information Measurement . . . 135
4.4 DISCUSSION & RESULTS . . . 138

4.4.1 Databases of image stimuli 140
4.4.2 Quantitative Comparisons of Saliency Methods 141
4.4.3 Qualitative Comparison of Saliency Methods 148
4.5 CONCLUSION 153

5 Multi-scale Discriminant Saliency (MDIS) 155
5.1 INTRODUCTION 155
5.2 DISCRIMINANT SALIENCY (DIS) CONCEPT 157
5.2.1 Centre & Surround Operation in Discriminant Saliency 157
5.2.2 Discriminant Saliency Formulation 159
5.3 MULTISCALE DISCRIMINANT SALIENCY (MDIS) MODELLING . . . 162
5.3.1 Multi-scale Classification Windows 162
5.3.2 Multi-scale Statistical Model 164
5.3.3 Multi-scale Statistical Learning 169
5.3.4 Multiscale Likelihood Computation 171
5.3.5 Multi-scale Maximum a Posterior 172
5.3.6 Multiscale Discriminant Saliency 174
5.4 DISCUSSION & RESULT 175
5.4.1 Quantitative Evaluation 178
5.4.2 Qualitative Evaluation 187
5.5 CONCLUSION 193

6 Conclusion and Future Works 195

References 200

List of Figures

2.1 General architecture of the ITT model [1] 33

2.2 PFT method demonstration on 1-D pulse and periodic signals [2] 37

2.3 A abstract picture of an ink bottle on a desk corner [3] 42

2.4 Evaluation Test 2 - Moving patterns 44

2.5 52

2.6 **Left** image with gray-level features, and **Right** image with orientation features are used for nearly uniform PDF 52

2.7 Example that highlight distinction between self-information and local entropy 54

2.8 6 dominant colours used in traffic signs 62

2.9 Modified architecture of a model:Saliency-based visual attention for Traffic Sign Detection 63

2.10 Yellow Colors Distribution in (a, b) space of CIELab 65

2.11 Traffic Sign Colour Distribution and Classification. (x_1, x_2) is another representation of (a,b) 66

2.12 Sample Feature Map 67

2.13 Sample Image 1 69

2.14 Sample Image 2 69

2.15 Sample Image 3 69

2.16 Sample Image 4 69

2.17 Sample Image 5 69

2.18 Phase Spectrum of Frequency-based Saliency Methods 71

2.19 PFT Lane-mark Extraction 72

2.20 Result 4 . . . 72
2.21 PFT Visual Saliency Method . . . 73
2.22 PQFT Visual Saliency Method . . . 74

3.1 Medium Band Filter Flow Chart . . . 79
3.2 Effective Saliency Region . . . 80
3.3 Samples and Saliency Maps of ITT, GBVS, PFT, and PQFT, and ENT methods (left-to-right top-to-bottom order). Please see the supplementary materials for sample sections of the video containing the original scene with the eye fixation marker overlayed on it and saliency maps of various methods. . . . 84
3.4 Standard Deviation of Each Sub-band in Camera-man Image . . . 85
3.5 Surrounding neighbourhood connectivity . . . 86
3.6 Energy Distribution according to PCA basis . . . 87
3.7 Spatial Saliency Maps by CON and KLD with (the bottom row) / without (the top row) DEC (decorrelation) . . . 88
3.8 Temporal Saliency Map without / with correlation . . . 90
3.9 Spatial Saliency Map: (left) without denoising, (right) with denoising . . . 92
3.10 Spatial Saliency Map: (left) without denoising, (right) with denoising . . . 93
3.11 Saliency Maps of Sample Images in Bruce' database [4] . . . 94
3.12 ROC and Inter-subject ROC curves on Bruce's database . . . 96
3.13 Saliency Maps of Sample Images in Judd's database [5] . . . 97
3.14 ROC curves on Judd's database . . . 98
3.15 Attention maps generate by six different methods . . . 104
3.16 Samples and Saliency Maps of ITT, GBVS, PFT, and PQFT, and ENT methods (left-to-right top-to-bottom order). Please see the supplementary materials for sample sections of the video containing the original scene with the eye fixation marker overlayed on it and saliency maps of various methods. . . . 106
3.17 MSDR Data Samle Image, Ground Truth, Importance Map . . . 106
3.18 Normalized Cross Correlation of Importance Map and Saliency Map . . . 109

4.1 Wavelet analysis of a sample image . . . 121
4.2 Wavelet (solid line) & Wavelet Packet Decomposition (solid and dash lines) for 1-D signal . . . 122
4.3 Illustration of DTCWT . . . 124
4.4 Illustration of dual-tree complex wavelet transform . . . 127
4.5 Inter-band and intra-band correlation of wavelet coefficients [6] . . . 131
4.6 Neil Bruce's database . . . 140
4.7 Kootstra's database . . . 141
4.8 ROC Curve of Neil Bruce's Database - WSS . . . 143
4.9 ROC Curve of Neil Bruce's database - DIS . . . 144
4.10 ROC Curve of Kootstra's database - WSS . . . 146
4.11 ROC Curve of Kootstra's database - DIS . . . 147
4.12 Saliency Maps 1 . . . 149
4.13 Saliency Maps 2 . . . 149
4.14 Saliency Map 3 . . . 150
4.15 Saliency Map 4 . . . 150
4.16 Saliency Map 5 . . . 151
4.17 Saliency Map 6 . . . 152
4.18 Saliency Map 7 . . . 152
4.19 Saliency Map 8 . . . 153

5.1 (L) : central contex (explicit manifold). (C): original image. (R): surrounding context (implicit manifold) . . . 158
5.2 Quad-Tree and Dyadic Wavelet Structures . . . 163
5.3 (U/T/V)HMT - MDIS - (IS)ROC . . . 180
5.4 Performances of (U/T/V)HMT-MDISin AUC,NSS,LCC,TIME . . . 184
5.5 Summary of all MDIS against AIM/DIS . . . 185
5.6 Saliency Maps 1A . . . 188
5.7 Saliency Maps 1B . . . 189
5.8 Saliency Maps 2A . . . 190

5.9 Saliency Maps 2B 191

5.10 Saliency Maps 3A 192

6.1 Relations and Overlapping between proposed methods ENT, WSS and MDIS 197

List of Tables

1 ENT,WSS,MDIS's quantitative results on N.Bruce's database iii

2.1 Main Colors of Traffic Sign in RGB[7] 63
2.2 Major Traffic Sign Colour Classification Parameters 67
2.3 Simulation Results (N_s : $detection - rate$) 68

3.1 Bias ratio vs. patch size 82
3.2 Evaluation results of Saliency Methods on Bruce's database 96
3.3 Evaluation results of Saliency Methods on Judd's database [5] 98
3.4 Normalized Saliency Value 100
3.5 Chance Adjusted Saliency 102
3.6 Normalized Saliency Values in various driving contexts 103
3.7 Difference in Normalized Saliency Values between go_alone and others 103
3.8 Importance order of MSRD object classes 107
3.9 Statistical data of Normalized Cross Correlation over video sequence 01 108
3.10 Statistical data of Normalized Cross Correlation over video sequence 02 108
3.11 Statistical data of Normalized Cross Correlation over video sequence 03 108

4.1 Wavelet Levels vs Frequency Range 116
4.2 Quantitative Result of Neil Bruce's Database- WSS 143
4.3 Quantitative Result of Neil Bruce's database - DIS 144
4.4 Quantitative Result of Kootstra's database - WSS 146
4.5 Quantitative Result of Kootstra's database - DIS 147

5.1 Comparisons of Evaluation Results for MDIS, DIS and AIM 181

5.2 UHMT - MDIS - DATA 182
5.3 THMT - MDIS - DATA 182
5.4 VHMT - MDIS - DATA 182

List of Abbreviations

PDF Probability Distribution Function

BY Blue-Yellow colour

H Shanon entropy

I image intensity

M Motion feature

RG Red-Green colour

ADAS Advanced Driving Assistance System

AIM An Information Maximization

AUC Area Under Curve

BSS Bayesian Surprise Saliency

BSS Bayesian Surprises Saliency

CAS Chance Adjusted Saliency

CON Conditional Entropy

CWT Continuous Wavelet Transform

DCT Discrete Cosine Transform

DEC PCA Decorrelation and Dimensionality Reduction

DIS Discriminant Information Saliency

DTCWT Dual-Tree Complex Wavelet Transform

DWPT Discrete Wavelet Packet Transform

DWPTBB Discrete Best Basis Wavelet Packet Transform

DWT Discrete Wavelet Transform

EM Estimation & Maximization

ENT Entropy-based Saliency

ENT-CON Entropy-based Saliency Method with Conditional Entropy

ENT-KLD Entropy-based Saliency Method with Kull-back Leilber Divergence

FIR orthonormal finite-infinite-response

FIT Feature Integration Theory

FPR False Positive Rate

FT Fourier Transform

GBVS Graph-based Visual Saliency Method

GD Gaussian Distribution

GGD Generalized Gaussian Distribution

GMM Gaussian Mixture Model

GRY Gray-scale Images

HMT Hidden Markov Tree

IM Importance Map

ISROC Inter-Subject Receiver Operating Characteristics

ITT Itti Saliency Model

KDPEE k-d tree entropy estimation

KL Kullback-Leibler

KLD Kullback-Leibler Divergence

LCC Linear Correlation Coefficient

MAP Maximum a Posterior Probability

MDIS Multiscale Discriminant Saliency

MIT a saliency method of Judd *et al.*in MIT [5]

MLL Maximum Likelihood

MPA Maximum a Posterior Probabilility

MSF Medium Sub-band Filter

MSRD Motion-based Segmentation and Recognition dataset

NORMXCORR Normalized Cross Validation

NP Non-deterministic Polynomial-time hard

NSS Normalized Scanpath Saliency

NSV Normalized Saliency Value

PCA Principal Component Analysis

PFT Phase Fourier Transform

PQFT Phase Quaternion Fourier Transform

PSS Pixel-based Scale Saliency

QFT Quaternion Fourier Transform

QWPTBB Quaternion Wavelet Packet Transform with Best Basis

QWT Quaternion Wavelet Transform

ROC Receiver Operating Characteristics

ROI Region of Interest

SR Spectral Residual

SSM Spatial Saliency Map

SSM Spatial Saliency Map

STFT Short-Time Fourier Transform

STSM Spatio-Temporal Saliency Method

T Thresholds

THMT-MDIS Trained Hidden Markove Tree for Multiscale Discriminant Saliency

TPR True Positive Rate

TSM Temporal Saliency Map

UHMT-MDIS Universal Hidden Markove Tree for Multiscale Discriminant Saliency

VHMT-MDIS Vector Hidden Markove Tree for Multiscale Discriminant Saliency

WSS Wavelet-based Scale Saliency

WT Wavelet Transform

WTA winner-take-all network

List of Mathematical Symbols

F Fourier Transform from spatial domain to spectral domain.

F^{-1} Inverse Fourier Transform from spectral domain to spatial domain.

$h(c)$ HUE difference [8] of a color c defined by Equation 2.5.1.

$\mathcal{HT}$ Hilbert Transform produces harmonic conjugate of a signal.

$\mathbb{Z}$ a whole natural integer number $[-\infty \ldots -1, 0, 1 \ldots +\infty]$.

π ratio of circumference of circle to its diameter.

$q = a + bi + cj + dk$ quaternion number q with components (a,b,c,d).

$\mathbb{R}$ include both integer number $\mathbb{Z}$ and fractional numbers .i.e 1.278.

$\ominus$ operator for computing differences across consecutive scales.

$\mid a \mid$ absolute of a number a.

$\triangleq$ define as.

δ partial derivative.

e Euler's number, approximately equal to 2.718 or a descriptor of wavelet-based scale saliency (WSS).

E_X statistical expectation of a variable X.

$(A \mid B)$ given operator $\mid$ describes dependence of event A on event B in conditional probability.

$\int$ integration of a continuous variable.

$K(P_A, P_B)$ Kullback-Leibler divergence of two probabilities P_A and P_B.

log natural logarithm operator with Euler's number exp as base with power a.

$MI(A, B)$ a measurement of mutual information between events A,.

$\neq$ not equal.

$\oplus$ operator for summing across consecutive scales or states.

$\hat{H}$ an approximation of entropy.

$\mathbf{\Sigma}$ a covariance matrix of a multivariate distribution.

$\mathcal{L}$ signal energy defined by Friston [9].

$H(A|B)$ entropy of conditional probability $p(A|B)$ of occurred event A given event B.

$\mathbf{H}$ a matrix of an entropy H.

$\mathcal{H}$ negative entropy used in calculation of free energy $\mathcal{F}$.

$\mathcal{F}$ a free energy defined by Friston [9].

$\mathbf{K}$ a matrix of Kullback-Leibler divergence.

μ_x mean value of a variable x.

$\mathbf{MI}$ a matrix of mutual information MI.

Θ parameters of a model.

$\mathbf{\Theta}$ a set of parameters of a model.

ϕ phase component of a complex number in Euler's form or a scaling function in wavelet-analysis.

ρ energy density of a descriptor.

$I(A)$ self-information of an event A as $I(A) = -\log(p(A))$ with its probability $p(A)$.

σ_x a variance of a variable x.

$\vec{x}$ a vector of spatial coordinates.

ψ a mother wavelet of wavelet analysis.

CHAPTER 1

General Introduction

1.1 MOTIVATION

Attention phenomenon is defined by neuro-psychologists as a neural system with the sole purposes of selecting meaningful and attractive information for the visual, auditory and motor systems [10]. It is an anatomical network of multiple cortex on which tasks are processed distributively. With flexibility in network formation, attention mechanism of human visions has been deployed for unevenly sampling data of the visual system in order to utilise its computational resources efficiently [11]. Exact roles of attention have been thoroughly discussed in several fundamental works by Yarbus [12], Neisser [13], Richards and Kaufman [14] and many other researchers. As a result, biological visual attention mechanisms are widely agreed to have at least basic independent but inter-related functions [15] as follows.

1. Selecting a region of interest (ROI) in the visual field
2. Choosing a feature dimension and value of interest
3. Controlling information flow through anatomical network of cortex in the biological visual system
4. Shifting from a ROI to the next one in time

Understanding dynamics of human vision systems and predicting visual attention remain formidable challenges. To tackle the latter challenge, this thesis focuses on studying computational saliency algorithms, which model first two among above four functions of biological visual attention. Saliency algorithms aim at measuring "interest" (salience) degree of different regions and features in arbitrary visual scenes; then, isolating and ordering those regions according to their decrementing order. Assumed

that dynamics of cortical networks are regulated by well-developed information theory [16, 17], informative quantities such as entropy, self-information [18], Kullback-Leibler divergence [19], etc are suitable for quantifying information from biological visual systems. Hence, the degree of interest in visual attention is inferred from how big those information quantities are. The more amount of information data contends, the more degree of interests or saliency these data have. Any computational model, which embraces this principal idea, can be classified as information-based saliency and its underlying mechanism is measuring interest / saliency as an amount of contended information. The rationale behind the approaches is biologically plausible as the human vision systems would focus on their processing capabilities on areas of complex textures and details (informative regions) instead of plain or empty regions. It is noted that similar concepts are exercised in the image compression approach where information theory plays an important role. Interpretations of saliency measurement with information theory inspire the development of proposed mathematical models and computational algorithms of information-based visual saliency in this thesis.

1.2 SCOPE

In information theory, the term "curse of dimensionality" refers to a phenomenon in which data are arranged in high-dimensional spaces (hundred-thousands of dimensions). In this phenomenon, data are unable be represented in low dimensional spaces and intuitively computed by basic statistical tools. Furthermore, increment in dimensionality leads to exponential expansion of data volume until available data becomes too sparse for estimating any statistically significant information quantity. To overcome this curse is required exponential increment of data samples which are unlikely available in case of visual saliency map computation. For example, an information quantity of 30x30 image needs estimating on 900-dimensionality (900-D) data space, and the estimation needs at least 2^{900} samples. It is almost impossible to collect enough samples with similar statistical characteristics. Moreover, such solution comes along with an intractable computational requirement for any information-based saliency solutions. Besides, the "curse of dimensionality" also hinders extension of the saliency methods to account for related issues such as attention in spatio-temporal domain and multi-scale visual attention. Any naive extension causes unimaginable explosion of data dimensionality, sample requirement, and computing time which is unable to be coped with. As performance of this estimation is vital to a information-related system; therefore, breaking the barrier, named as the "curse of dimensionality", leads to achievement of statistically accurate information-based saliency.

1.3 OBJECTIVE

The previous section 1.2 discusses how badly the "curse of dimensionality" affects performances of general information-based saliency methods. Therefore, the main objective of this thesis is investigating multiple solutions and computational models for a computationally feasible information-based visual saliency method. To achieve this objective, following directions are worth investigating.

1. To investigate information measurement concepts such as conditional entropy, self-information, Kullback-Leibler divergence as units of saliency.
2. To formulating saliency in Decision theory framework by discriminating interested regions from the rest of images.
3. To develop more efficient algorithms for estimating high-order statistics of visual data instead of the traditional marginal probability distribution of pixel-intensity histograms
4. To relate multiple characteristics of human vision systems and information-based saliency frameworks.
5. To investigate how multi-scale hierarchical nature are fitted in information-based saliency approaches and how prior knowledge affects attention.
6. To develop both spatial and spatio-temporal data as attention is surely attracted to moving objects.
7. To verify that information-based saliency maps is correlated with ground-truth of eye fixations provided by human subjects.

1.4 CONTRIBUTION

As the previous section 1.3 states that the main objective of this thesis is studying information-theoretic based saliency methods, this chapter briefly describes three main contributions.

1. The Entropy-based Saliency Method (ENT) is
   - reducing data dimensionality by medium spatial frequency (MSF) filters [20]
   - adapting kd-tree entropy estimation method (KDPEE) [21] for fast saliency computation

- extended into spatio-temporal domain for video processing
- evaluated quantitatively and qualitatively with standard image databases (Bruce's [4] and Judd's [5] databases)and reliable video databases.

2. The Wavelet-based Scale Saliency (WSS) is

- proposing wavelet sub-band energy descriptors in place of pixel-based ones,
- adapting dyadic scale framework into the scale-saliency [22],
- developing a suitable mathematical formula for estimating WSS saliency,
- evaluated quantitatively and qualitatively with standard image databases (Bruce's [4] and Koostra's [23] databases).

3. Multi-scale Discriminant Saliency (MDIS) is

- interpreting centre-surround operations as multi-scale binary classification of central and surrounding features,
- adapting MDIS principles into Hidden Markov Tree (HMT) framework,
- developing three derivatives of MDIS, UHMT-MDIS , THMT-MDIS , VHMT-MDIS with respect to how HMT parameters are learned.
- Being evaluated quantitatively and qualitatively with a standard image database (Bruce's [4] database).

1.5 ORGANIZATION

This whole thesis is devoted to describing proposed solutions for the curse-of-dimensionality problem, as mentioned in Section 1.2, to achieve fast and efficient information-based saliency algorithms 1.3. As the thesis mainly focuses on describing and explaining about three main contributions, it is organised into three main Chapters 3,4, and 5 accompanied by Chapter 2 for a literature view, and Chapter 6 as a conclusion.

Chapter 2 reviews current literatures of biological visual attention phenomena as well as existing computational saliency algorithms. As these two topics are closely related and results of saliency computation quantify degree of visual attention at specific locations, it is fruitful to discuss relation between saliency and eye-fixation concepts and available saliency algorithms in computer vision literature. These models can be classified into three main categories

1. feature-based saliency methods

2. frequency-based saliency methods
3. information-based saliency methods

according to how they simulate centre-surround operations for computation of saliency values. Among three categories, information-based models are further elaborated with conceptual thought experiments by Attneave [3] and common mathematical theories as well as their pros and cons for evaluating information quantities as saliency values.

After the literature view in the Chapter 2, Chapters 3,4,5 describe our main works ENT, WSS, and MDIS in details, which are purposely arranged in accordance with increasing order of sophistication and robustness.

In Chapter 3, ENT model demonstrates how unnecessary data can be removed by medium-spatial-frequency filters [20], how information quantity can be estimated by the fast KDPEE technique [21]. Moreover, it also describes how ENT is extended into spatio-temporal domain for processing video data.

Then, Chapter 4 recommends alternative time-frequency descriptors with a modified mathematical model WSS which allows scale-saliency estimation from the marginal density distribution of the proposed descriptors. Besides describing rationales behind the wavelet-transform based descriptors and its importance in fast and efficient implementation of multi-dyadic scale saliency, this chapter also investigates effects of different wavelet transformation on saliency estimation.

Instead of modelling only marginal relativity like WSS, MDIS in Chapter 5 employs Hidden Markov Trees with multi-scale strategy for estimating saliency values by formulating discriminant power between centre and surrounding classes with their corresponding wavelet features. Moreover, it points out conceptual equivalence between MDIS and other information-based approaches as well as describes usage of different derivatives of Hidden Markov Tree (Universal (U), Trained (T), and Vector (V)) and their performances.

Finally, principle ideas and significant contributions of this thesis are once again emphasized in Chapter 6. Furthermore, limitations of proposed approaches are discussed as well as possible extensions are laid down for further research directions.

1.6 SUMMARY

In a nutshell, this chapter involves six main sections MOTIVATION, SCOPE, OBJECTIVES, ORGANIZATION and SUMMARY. Section 1.1, MOTIVATION, describes what

the attention phenomena are and how information-base saliency methods are their useful quantifying methods. However, Section 1.2 SCOPE mentions common dilemma of the saliency methods, the "curse-of-dimensionality", which prevents them from efficient implementations or extensions to different feature domains. Then, coming with main objectives and contributions of the thesis: ENT, WSS, and MDIS are Sections 1.3 OBJECTIVES, 1.4 CONTRIBUTION. Section 1.5, ORGANIZATION, lay out clear structures of how the works are organized in chapters of this thesis. Finally, this section SUMMARY 1.6 attempts to summarize flows of contents or arguments as well as mention possible future research directions.

CHAPTER 2

Literature Review

In spite of importance of visual system in daily life, very little is known about our visual perception and there exist many questions which have not been fully answered yet. For example, whether do human visual systems observe all information on the scene at once or receive a small portion of that whole information once at a time ?. Treisman [24] found supporting evidences for the later hypothesis in which only a portion of visual information reaches later processing stages and has impact on human being's conscious awareness at a time. The procedure of choosing appropriate portion of visual information is named as Visual Attention. It explains how a normal human brain can response to stimuli in real-time without exhausting all biological computational units. Studying and modelling such ability of prioritizing visual information for further processing greatly benefits the development of computer vision systems, which are also required to deal with large amount of data with only limited computing capabilities. The set of computational models inspired by visual attention strategy is generally named as computational visual saliency algorithm.

2.1 INTRODUCTION

A human vision system is able to cope with variety and changes of scenes due to its attention processing, a well-structured prioritization of incoming information. It selects only relevant stimuli from ocean of irrelevant and insignificant features, which plays a vital role in eliminating unnecessary computational loads in later stages. The receptive fields of human eyes are composed of multiple neural layers [25] which processes scenes from coarse to find scales. That arrangement of receptive fields is analogous to the popular pyramidal processes in computer vision [26], which is an excellent strategy for reducing the overall complexity of any systems since information can be localized

and processed at a suitable scale. However, that hierarchical structure also requires developments of additional mechanisms or regularization of how information might flow within their layers.

Pioneering in researches of computational mechanisms for visual attention, Tsotsos *et al.*[27] initialize basic principles to investigate the attention phenomenon in a hierarchical framework for visual input signals. Searching a scene with no particular targets is NP-complete problems, so is attention when general searches are carried out without specific objectives until interesting image elements are spotted. To reduce computational complexity of NP-hard problems, biological visual systems utilizes a pyramidal processing architecture, which are supported by neural cognitive researches [28–31]. Such pyramidal architectures are observed in primate brains in which coarse-to-fine processing layers are realized by increment of receptive fields and neuronal connectivity along visual pathways [32]. Though pyramidal processing may optimize computational performance for solving NP-complete problems such as attention, Tsotsos *et al.*points out four basic challenging scenarios: context , blurring, cross-talk and boundary effects arising in the pyramidal framework. In these contexts, responses of coarsest layers in pyramid may be traced back to input features at different spatial scales. Consequently, responses generated at a top layer according to not only its own features but also a variety of different surrounding context at a bottom layer. For example, blurring problems are caused by wide-spread unnecessary activation of neurons from fine-to-coarse, or bottom-to-top layers due to small localized events in pyramidal processing. Moreover, blurring prevents effective localization at the top-layer due to activities of redundant neurons along pyramid layers. In addition, cross-talk and boundary effects are waste products of unnecessary neural activities from two or more localized events at bottom-layers. Cross-talk usually occurs when distances between isolated image elements are relatively small and their spatial locations are around the central area of images. It causes undistinguished attention for two independent and isolated visual features. Then, boundary effects happen when responses of central image units are significant stronger than those of boundary units; in order words, it causes unjustified bias in visual interest of scenes. These effects are caused by asymmetric distribution of neural connections across layers when boundary neurons have fewer connection to upper or lower pyramidal levels than central ones.

The multi-layers' structure and problems of pyramidal visual attention theory apparently fits well into Marr's levels of analysis [33] about computer vision. Though Tsotsos [27] and Marr [33] have different research starting points, they accidentally come up with common sets of theories and problems. This coincidence inspires a combined the-

ory of how the biology scientist and the computer vision engineer view visual attention processes. The combined theory motivates generation of computational models for visual attention or computational visual attention. These combined models often give far more comprehensive mathematical and algorithmic description of biological mechanisms behind human vision than only biological or computer vision models alone possibly do. Compared to other approaches, their greatness in details have both pros and cons. The first advantage of such detailed models is a feasible simulation of how a computation system performs against the biological vision system as well as other models, providing similar images are used as input signals and presented to human subjects. Second advantage is focusing on biological connections with suitable computer vision algorithms instead of pure computational efforts. Moreover, such combined theories are falsifiable, which can make testable predictions against salient, new, not yet observed or counter-intuitive information. However, their first disadvantage is also the comprehensive specification since it allows more difficult objective evaluations which require carefully designed and experimental validation. The second disadvantage is an open definition of how attention should be represented in the brain. Although many evidences [34] [35] [36] support single representations or layers for the whole observed field of the pyramidal framework, existence of such layers in the brain is still debatable due to selective nature of attention itself.

The unclear definition of representation for attention does not stop a quest towards the best single representation, called saliency maps, for computational attention. Saliency concept and definition are discussed in Section 2.2. As saliency is the main current stream of visual attention research, there are number of models such as Clark and Ferrier [37], Sandon [38], Itti *et al.*[1, 36], Walther [8], Gao *et al.*[39], Bruce *et al.*[4] in the last decade. The mentioned methods follow common steps:

1. Visual stimuli generate conspicuity within particular feature dimensions by a centre-surround operation which projects visual features into a single and normalized unit.

2. Interests are then determined by a degree of differences between features at the location and its surrounding.

3. Those interests are combined into a global measure of saliency maps and evaluated with eye-fixation ground-truth.

These features maps, conspicuity maps and regions of interests are popular due to its coexistence in both computational and biological aspects of attention. The feature

maps correspond to early representations of several stimulus with different characteristics. The relation between these visual stimuli and eye movements in regions of interests are further discussed in Sub-Section 2.3.1. Moreover, Sub-Section 2.3.2 provides brief overview how simulations should be carried out with proper evaluation to justify performance of saliency methods. Sub-Sections 2.3.1 and Sub-Sections 2.3.2 in Section 2.3 cover both input features and output evaluation methods for modelling computational saliency as a signal processing problem. Then, the main computational models of visual saliency are categorized in three categories: feature-based saliency, frequency-based saliency and information-based saliency in Section 2.4. While each category is discussed about how it formulates interest / saliency measurement from input visual stimuli, information-based saliency is especially focused since it motivates proposed models in this thesis. Besides overview of saliency concepts and models, Section 2.5 describes two applications in which visual saliency directly provides solutions for practical computer vision problem rather than pre-prepocess of choosing interest points or regions. It shows usefulness of saliency algorithms in larger sets of vision applications rather than only detection of interest points. Finally, Section 2.6 summarizes significant contents of this chapter.

2.2 SALIENCY CONCEPT

As shown in the title of this thesis, its main concern is modelling saliency, which is briefly mentioned in the last section. Due to its importance to computational modelling of selective attention as well as the subject matter of this thesis, more discussion are required to elaborate the saliency concept. According to the Merriam-Webster dictionary, the term saliency is used for describing the state of being "salient, standing out conspicuously, prominent, and notable significance". In terms of vision research, saliency defines orders of interests according to which corresponding features should be captured, and forwarded to higher processing levels. Moreover, the selection of features needs to be intuitively consistent in attracting human gaze regardless of any assigned tasks in mind since the spotlight of attention is driven toward interesting locations with the maximum saliency value [24]. Though saliency concept often appears along with saliency maps in several models [1] [40] [18], the saliency concept is more general than saliency maps are. In the models, saliency is a unit for measuring degrees of interests while saliency maps are just one among several methods of utilising this measurement. For example, computational attentions can be classified into four main branches: selective routing, saliency maps, temporal tagging and emergent attention [41], all of which are related to saliency unit. In other words, saliency maps always

need salience notion as units for measurements; however, saliency concept still exists in other approaches where saliency maps are not formed.

It is worthy mentioning about how covert and overt eye movements are disregarded in computing saliency values and topographical maps. There has been no simple means of tracking the focus of convert attention; therefore, the prediction of eye movement patterns are studied in many models regardless of these eye movements. Though there is a possible but complex mean of observing covert focus of attention [42] and such a solution can be used for tracking covert focuses of attention in order to access model performance, it is implicitly assumed that overt and covert eye movements have similar effects to simplify experimental processes. Moreover, overt attention will be assumed by default, otherwise is specially mentioned. These assumptions provide suitable platforms for discussing saliency with other related issues in the rest of this chapter.

2.3 SALIENCY MODELLING

The previous Section 2.2 defines an abstract definition about saliency as a unit for measuring visual interest. To measure this measurement quantitatively, signal processing models needs to be developed and evaluated. Before formulation of any computer vision model, two questions need to be answered first.

1. What are good visual features & stimuli for an application ?
2. What experiments and evaluations need to be done ?

As saliency concept is inspired by attention phenomenon of human visual system, both input features and evaluations have to be closely related to this biological process. Therefore, the two questions above can be rephrased for saliency modelling purpose as follows.

1. What are good visual features for attention ? How are these stimuli related to eye movement ?
2. What experiments need to be done for evaluating saliency maps against eye fixation maps, a biological ground truth?

These questions are discussed in following Sub-Sections 2.3.1 and 2.3.2 accordingly.

2.3.1 Visual Stimuli & Eye Movement

Correlation between visual stimuli and eye movements is important in the context of visual attention. To study that relationship, Buswell [43] has carried out one of early influential experiments into reading and picture-viewing by non-intrusive eye tracker. In this experiment, the tasks are carried out without any specific searching target; therefore, eye movements of subjects are completely driven by bottom-up features. Yarbus studies effect of top-down visual search tasks on eye movements by conducting eye-tracking experiments with different objectives [12]. According to results of two experiments, observers tend to focus on particular parts of the scene regardless of specific searching assignments; it confirms that early bottom-up stimulus natively directs dynamic saccades of eye movements. In other words, biological saliency maps can be formulated by mapping all eye-fixation points on a canvas of visually perceptual fields in theory. In practice, a Gaussian blur effect is applied on such canvas to ensure a smooth and intuitive saliency representation. Generated artificial saliency maps, computed from these early bottom-up stimuli, can be evaluated by its correlation with ground-truth maps. The study of Privitera [44] points out varying degrees of correlation between basic features such as edge, curvature and deployment of eye movements and attention. Many visual features, which are highly correlated with eye fixations, are outlined and shortly discussed. It helps identify which features have mostly influenced where eye-saccades are deployed. Privitera and Stark have initialized their exploratory effort on this matter by considering the relation of 10 different local image contents and deployment of eye movements.

1. Edges per unit area with Canny edge detection algorithm [45]
2. Masks selective for high curvature
3. Positive centre - negative surround in a 7x7 square window.
4. Gray-level orientation differences in Gabor masks like [46]
5. Local symmetry measure
6. Michaelson contrast: $C = \mid (L_m - O_m)/(L_m + O_m) \mid$ where L_M is the mean luminance of a local 7x7 region and O_m is the overall mean luminance of the whole image .
7. Texture variance measure by an entropy value: $N = \sum_{i \in G} f_i log f_i$ where f_i is from gray-level histogram in the neighborhood G.

8. High frequency coefficients of the Discrete Cosine Transform.

9. Centre-Surround by Laplacian of Gaussian operator

Each of above 10 operators has varying degrees of correlation with human fixation points with the exception that the Discrete-Cosine-Transform (DCT) based operator was uniformly poor. A common drawback of all above local visual features is their inconsistency in correlation with eye fixations. That is, each of them performs well in a subset of images but very poorly in others. It seems that none of the operators are good enough to predict eye movements universally for all types of images. Moreover, the inconsistency of local features means that they alone are not sufficient for good prediction of eye fixations. itemize Tatler, Baddeley and Gilchrist [47] conduct similar but also important study in which correlation of contrast, edge and chromaticity at various scale spaces and eye fixation points are examined. Perhaps, their most significant finding is that models of neurons in networks of visual cortex inspire many operators based on combinations of angular and radial frequency, contrast and colours. It confirms similarities between various feature measurement of stimuli and responses of different neuron types, akin to those appearing primary visual cortex of primates. These neuro-biological plausible feature measurement shows good predictive properties of fixation patterns. Among those features, edge content and contrast primarily dominate accurate prediction of fixation points, especially in high spatial frequency. In addition to discovering eye-fixation predictive properties of many features, Tatler *et al.*'s work also focuses on problems in methodology of assessing agreement between features and eye fixation data, which closely relates to accuracy of many vital experiments in this subject. The importance of this matter with respect to quantitative assessment in this dissertation warrants further discussion in later chapters.

Pomplun [48] leads another study into correlation with eye movement deployment. Different from two prior experiments [44][47], Pomplun considers how top-down information can affect distribution of eye fixations. High-level tasks are commonly related to searching operations in which a particular pattern needs to be found in complex natural scenes. In the experiment, fixated points and regions are first collected and then compared with the target's patterns of intensities, contrast, spatial frequency and orientation. In consideration of each above visual features, statistically significant correlation of feature measurement is found between target items and fixated regions. Among local features investigated in the Pomplun [48] experiment, intensity and contrast have the strongest agreement with fixation regions. Moreover, eye movements of observers in task-related experiments would be partially influenced by specific goals of visual search, there is always an inherent cap on performance of predicted fixation re-

gions over eye-movement deployment. To put differently, the relation of basic features and target patterns does not overwhelmingly bias eye-movements, and bottom-up feature based selection does remain its influence in certain extent despite in visual search tasks.

Although Tatler *et al.*and Pomplun *et al.*'s work [47, 48] reveal few clues for answering the question "what" nature of stimuli tends to draw eye gaze of observers, the experimental setup and results are not comprehensive due to ignorance of relation between visual features and their possible neural representation. In fact, some image operators are predictive of fixation patterns but useful for understanding what happens in the brain. In contrast, cortical processes or principled operators are inspired by hypotheses which have been tested by psychologist, neuroscientists to confirm effectiveness in predicting fixation regions. However, their computational implementation does not generate up-to-expected predicting agreement with eye fixations. In fact to date, there is no simple operators found for a reliable prediction of gaze patterns at all images. Some image channels correlate with these patterns of eye movement for a subset of samples at some of the time, but no particularly explicit conclusion can be made about the best feature. Experimental results indicate that no particular visual features or image operators as dominant underlying neural representation.

2.3.2 Experimental Setup & Data Evaluation

Experimental Setup The problems are not only with visual operators but their evaluation against fixation patterns as well due to difficulties in experimental designs. The experiment needs carrying out with following constrains to guarantee the statistical reliability of its results.

- It starts with the number of image samples with eye fixation data. In some cases, the set of images is sufficiently small; hence, it is ambiguous how well that small number of samples can present general situations. Assuming a large collection of images are used in the experiments, the respective eye fixation data are enough to sufficiently approximate human responses of natural scenes.

- Moreover, there is even bigger issue lying in how observers are instructed during experiments to disentangle top-down and bottom-up influences. For example, observers may have strong prior on contents of scenes if majority of images are taken in outdoor environment, sunny days, which makes them tend to fixate on certain spatial locations or features of the scenes. While choices of familiar natural scenes might cause significant bias in fixation distribution, selection of unpopular

stimuli has a problem on its own as well.

- For example, images consist of such stimuli as painting, terrain and landscape photographs may eliminate prior bias from observers; however, these stimuli in many cases have very different spectral profile than natural scenes. Consequently, behaviours of eye movements in those cases can not be generalized to fixation patterns in natural viewing conditions owing to unnatural neural response from unusual contents. As methodology of experimental conduct and design is vital to overall performance evaluation in the subject matter, further discussion about how to produce natural stimuli with minimum prior knowledge can be found in later chapters.

- An additional issue pertains to the experimental methodology is assessing performances of proposed methods. Fundamentally, the evaluation process selects a number of discrete coordinates for artificial fixation points and correlate them with human eye-fixations as ground-truth data according to presumed metrics. For selection procedures, variety of solutions from hierarchical Winner-Take-All (WTA) networks to clustering algorithms are utilized; for example, the simplest approach is determining the extent to which a few largest peaks correspond to human fixation points. However, this method misses the assessment of saliency regions away from those peaks. In addition, there is also ambiguity of how many peaks from saliency representation would be chosen and which cut-off threshold value would be useful for segmenting regions predictive of eye-movements from the rest of visual scenes. Many of unknown parameters leads to the diversity of assessing algorithms which then make it difficult to gauge a single quality of underlying saliency representation. In general, assessment of saliency models depends on both the quality of underlying saliency representation and choices of numerical evaluation methods.

Data Evaluation To make performance evaluation and comparison of computational saliency approaches reliable, it is necessary to utilise standardized sets of evaluation tools. Therefore, several evaluations methods such as Receiver Operating Characteristics (ROC) [47], Area Under Curve (AUC) [47], Inter-Subject Receiver Operating Characteristics (ISROC) [49], Normalized Scanpath Saleincy (NSS) [50] [51] and Linear Correlation Coefficient (LCC) are proposed for fair comparison between different saliency methods in the literature. These evaluation methods are summarized in the rest of this section. It is noted that all evaluation results are produced for image databases or sets of images unless otherwise stated.

ROC Tatler *et al.*[47] proposes usage of Receiver Operating Characteristics (ROC) curves as solutions for comparing the quality of saliency measurement. For a binary classifier, a ROC curve plots a hit rate or True Positive Rate (TPR) as a function of a false alarm rate or False Positive Rate (FPR) , $TPR = f_{ROC}(FPR)$. For saliency evaluation with eye-fixation ground truth, a ROC is constructed as follows.

1. Numerous thresholds are uniformly sampled from ranges of saliency measurement.
2. Each threshold is set for a binary classifier to distinguish between fixated and non-fixated locations; in other words, it generates artificial fixation data.
3. From collected and generated fixation data, hit rates (the ratio of correctly classifying fixated locations) and its false alarm rate (the ratio of misinterpreting non-fixation points as fixated data) can be found.

It is noted that thresholds (T) are linearly varied from the largest value (strictest criterion) with hit rate , false alarm rate {0,0} to the smallest value (easiest criterion) with $\{TPR, FPR\} = \{1, 1\}$. In between the two extreme points of the ROC, hit rates and false alarm rates are monotonically varied in the range [0,1], and higher hit rate value for lower false alarm value demonstrates desired pair of ROC values. In other words, the more the ROC curve moves toward top-left corner of plots whose TPR and FPR values are $1, 0$, the better a saliency method performs. For an image database, hit rates (TPR) and false-alarm rate (FPR) at each threshold T are computed from eye-fixations and saliency maps of all images in the database. Therefore, there is a single ROC plot for an image database with respect to each saliency method. All ROC plots in this thesis are generated for all images of a database rather than a single image.

AUC Though ROC helps compare saliency methods visually, it is necessary to justify the comparison numerically. Therefore, Tatler [47] also proposes that the overall performance of each saliency method can be quantified by the Area Under the ROC curve (AUC). In other words, AUC is an average performance of ROC over all possible thresholds. Advantages of AUC over ROC are

1. assessing saliency measurement over all possible thresholds
2. avoiding the problems with a single threshold or a specific selection mechanism
3. offering a general quantity of agreement between a measure of saliency and eye fixation of human data.

As ROC is plotted from saliency maps and eye-fixation of all images in a database, AUC values is also used for justifying performance of saliency methods for the whole database instead of a single image.

ISROC ROC curves and AUC values are only useful for comparing different computational saliency approaches, but they don't emphasize comparison between these methods with generalized human performance (correlation across subjects). It is necessary to have alternative evaluation methods for modelling such comparison. Therefore, the inter-subject ROC curves (ISROC) proposed by Harel *et al.*[49], is developed to describe performance of human visual system versus that of computational saliency methods. The ISROC curve is constructed as follows.

1. Fixation points of all but one human subjects are blurred with a circular, decaying kernel with decay constant matched to that of cone density in human retina. The blurred fixation map was treated as a saliency map derived directly from human fixations.

2. With fixations of the leaved-out subject, an AUC was computed for that saliency map.

3. Repeating the first two steps for all subjects and averaging all generated AUC values gives a subject-based AUC for that image.

4. Repeating the first three steps for all images in a database gives subject-based AUCs for all images.

5. Sorting the above AUCs values incrementally shows how well human subjects predict fixation points of the other subjects for each image or how difficult the image is for human visual attentions.

6. For a computational saliency algorithm, saliency maps are generated for all images, so are their algorithm-based AUC values.

7. The ISROC curve plots algorithm-based AUC values as a function of incremental subject-based AUC values.

If an ISROC plot of a saliency method is above the main diagonal, this saliency method performs better than an average human in predictions of eye-fixations. If an ISROC plot of a saliency method is above that of another method, the first performs than the latter regardless of how difficult testing samples are.

NSS is an average of saliency values from a normalized saliency maps with standard deviation at locations of eye positions. Eye positions include both eye-fixation points where eyes stand still and focus and saccade points where eyes just pass through when moving from one fixation point to another. NSS is computed as follows.

1. Firstly, all saliency maps are normalized with unit standard deviation.
2. Secondly, normalized salienchy values are extracted at locations of all eye-fixation along provided a scan-path of human eyes.
3. Thirdly, NSS is an average value of these extracted saliency values.

When $NSS = 1$, saliency values at eye-fixation locations are one standard deviation above average. When $NSS \geq 1$, normalized saliency values at fixation points are significant higher than at other locations. Meanwhile, $NSS \leq 0$ means that saliency maps perform worse than randomly generated maps. For an image database, its NSS value is an average NSS values of all images.

LCC measures strength of linear relationship between saliency maps and eye-fixation maps and it is computed as follows.

1. Build a ground-truth saliency map (G) by convolving maps of eye-fixation with Gaussian kernels for smoothing
2. Compute covariance between the ground-truth saliency map (G) and saliency maps (S) $cov(G, S)$
3. Compute standard deviation of ground-truth maps (G) and saliency (S) as σ_G, σ_S
4. The LCC value is ratio between covariance of G,S over a product of their standard deviation, $\frac{cov(G,S)}{\sigma_G \sigma_S}$.

When CC reaches its extrema +1/-1, it shows a perfect linear relationship between G and S. For an image database, its LCC values is an average LCC values of all images.

2.4 COMPUTATIONAL MODELS

The previous Section 2.3 has discussed about suitable visual stimuli and evaluation processes for designing a computational saliency model. In this section, developments of the models over past decades is categorized and details of few essential model are

summarized. Recently, intensive efforts have been put in researches of visual attention modelling, especially computational saliency maps. Therefore, quite a number of saliency techniques are introduced, for example, Itti saliency model (ITT) [1], An Information Maximization Model (AIM) [18], Phase Fourier Transform (PFT) Saliency [52] etc. Though their formulation and implementation are varied, there are common characteristics of underlying mechanisms and similarities. Basic steps of any bottom-up saliency formulation are selecting low-level feature descriptors and designing algorithms for fusing these features into saliency maps. Hence, low-level descriptors and fusing mechanisms can be used for classification of saliency methods.

According to signal processing views, saliency techniques can be classified by three types of available descriptors: purely temporal, spectral, and hybrid temporal-spectral descriptors. In most of vision systems, pixel data are first low-level features, captured by cameras. Analogously, light-sensitive cells on human retinas capture biological pixel data for later processing. If human vision is solely considered as a signal processing system, other data representation in many different domains are feasible. For example, basic signal processing tool ,Fourier transform , converts image data from spatial domain to frequency domain. As fundamental tools in many signal processing algorithms, Fourier transform helps to represent data on pure frequency domain. Moreover, as image features can be extracted from different frequency range and human visual perception is rather sensitive to features' spectra, multiple spectral analysis is considered useful tools in variety of saliency techniques [53]. Besides pure time or frequency domain, hybrid time-frequency representation are popular choices as well, thanks to the rapid development of wavelet transform in the last decade [54]. Such transforms offer local rather than global frequency analysis [54]. More than just another alternative space, wavelet-based representation would intuitively be more successful as descriptors for saliency maps construction because spatial locations are as important as spectral information to the visual systems according to two-stream hypothesis [55].

Besides image descriptors, saliency techniques are categorized by their underlying computation of saliency values, more specifically, modelling of centre-surround operations. The operation are widely accepted as the main principle for saliency computation [24]. Furthermore, these operators can be interpreted in three basic ways: contrast-based mechanisms with pixel descriptors, phase-based mechanism with frequency descriptors and higher-order information-based mechanism with wavelet descriptors. In psychological mechanisms, centre-surround operators play basic roles in determination of saliency values for each location. The operator is deduced from researches about retina cells, whose responses become very strong after observing dark

regions on bright background or vice versa. Based on this retinal property, many early saliency methods [1],[35] have mimicked such phenomena by linear contrast of low-level image features across multiple scales with pixel data as chosen descriptors. The contrast operators are only suitable for modelling linear relation between pixel descriptors and saliency values, while their roles in other data domains are very limited. For instances, in frequency space, differences between central-surrounding frequencies do not benefit spatial saliency map construction due to the uncertainty principle [54]; thus, alternative operators are necessary. As frequency coefficients can be fully characterized by amplitudes and phases, each of them contains vital but non-overlapped spectral information; therefore, few methods have used these amplitudes [52] and phases for saliency computation. Moreover, contrasts of central pixels and their neighbours actually contain first-order linear operators only while useful information might be found in higher-order correlations. To overcome such shortcomings, alternative information-based approaches are introduced in this thesis in order to describe higher-order statistics between central and surrounding features such as conditional entropy, information divergence, etc. The development of statistical non-linear contrasts and other higher-order centre-surround operators greatly benefits accurate estimation of image data with different characteristics as well as helps avoid curse of dimension in non-linear statistical estimation of saliency values.

2.4.1 Feature-based Saliency Method

Saliency notions became popular in the literature of computational attention after introduction of saliency maps in Feature Integration Theory (FIT) [24]. Basically, the model depends on various elementary image channels; distinct features are subsequently integrated into a single topographical representation of salience (saliency map). The map helps to drive spotlights of attention by selecting a spatial location such that it coincides with the largest peak on this map. In the FIT context, saliency maps refer to results of combined pooling basic feature maps; meanwhile, a saliency value itself is an output of operation combining basic set of features.

In Treisman and Gelade's original description of FIT model, there was little specification regarding nature of signals; therefore, Koch and Ullman [35] introduced a FIT-like model but with three specific features: colour, intensity, orientation which are actually preferred by human visual systems. Moreover, a distinct contribution of Koch and Ullman's research is introducing a winner-take-all (WTA) network as a post-processing step for sequentially select attentive spots. Such WTA networks have found widespread usages even outside selective attention researches; for example, WTA is applied

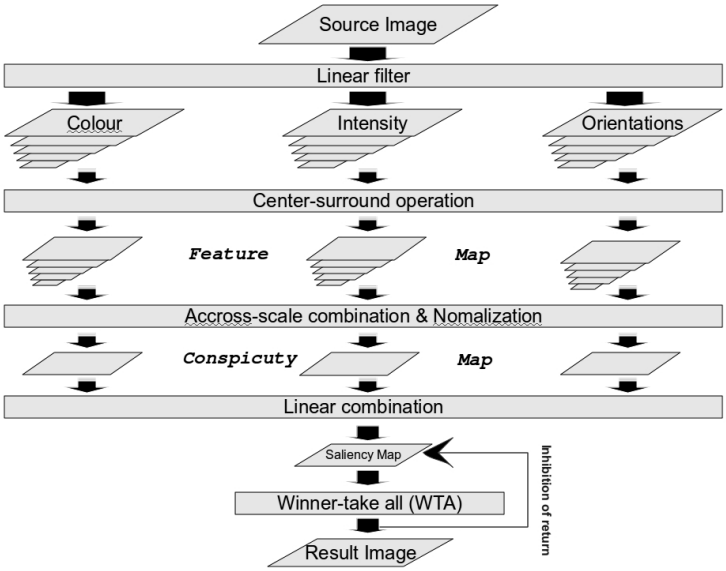


Figure 2.1: General architecture of the ITT model [1]

to the problem of modal control of a robot head by Clark and Ferrier [37]. Koch and Ullman's researches have provided theoretical FIT-like computational models with three basic features and a sophisticated WTA network; then, the real computational implementation with all mentioned details are released for research community by Itti *et al.*[1]. It provides specific feature definitions and claims concerning neurobiology and visual search performances. With detailed descriptions in their publication [1], [36] and reproducibility of simulation results, the particular approach has enjoyed considerable citations and chosen a de-facto referenced method. The figure 2.1 graphically demonstrates what steps are involved in the model.

The rest of this section briefly describes Itti *et al.*'s model as an image processing system. Provided that initial input signals are static colour images, they will be subsampled into nine spatial scales by using dyadic Gaussian pyramids. Then, they are progressively low-pass filtered and sub-sampled, which yields nine size-reduction images ranging from 1:1 (scale zero) to 1:256 (scale eight). The colour input images are analysed into different feature maps, which are in turn processed by a set of linear "centre-surround" operators, mimicking psychological behaviours of biological visual cortex. Typically, visual neurons are most sensitive in a small region of the visual space (the center), while stimuli presented in a broader region concentric with the centre (the surround) inhibit the neuronal response. Centre-surround operation is implemented in different feature maps as the difference between fine and coarse scales: the centre is a pixel at scale $c \in 2, 3, 4$, and the surround is the corresponding pixel at scale $s = c + \theta$ with $\theta \in 3, 4$. The across scale difference is represented by " $\ominus$ " which is obtained by

interpolation to the finer scale and point-by-point subtraction as follows.

$$I(c,s) = \mid I(c) \ominus I(s) \mid \tag{2.4.1}$$

$$RG(c,s) = \mid (R(c) - G(c)) \ominus (G(s) - R(s)) \mid \tag{2.4.2}$$

$$BY(c,s) = \mid (B(c) - B(s)) \ominus (Y(s) - B(s)) \mid \tag{2.4.3}$$

$$O(c,s,\theta) = \mid O(c,\theta) \ominus O(s,\theta) \mid \tag{2.4.4}$$

Those feature maps are later combined into three "conspicuity maps" for intensity, colour, and orientation at the scale ($\sigma = 4$). These maps are obtained through across-scale addition, " $\oplus$ ", which consists reduction of each map to scale four and point-to-point addition.

$$\overline{I} = \oplus_{c=2}^{4} \oplus_{s=c+3}^{c=4} N(I(c,s)) \tag{2.4.5}$$

$$\overline{C} = \oplus_{c=2}^{4} \oplus_{s=c+3}^{c=4} [N(RG(c,s)) + N(BY(c,s))] \tag{2.4.6}$$

$$\overline{O} = \sum_{\theta \in 0,45,90,135} N(\oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} N(O(c,s,\theta))) \tag{2.4.7}$$

The last step of this whole process is combining those conspicuty maps together into the final saliency map.

$$S = \frac{1}{3}(N(\overline{I}) + N(\overline{N}) + N(\overline{O})) \tag{2.4.8}$$

At first, Itti *et al.*'s model [1] only operates on three main channels: intensity, colour and orientation; however, many additional features are constituted into the framework. Then, saliency map becomes unified representation of original along with extra features such as depth [56], motion [57, 58], symmetry [59], faces [57], or skin [58].

As Itti *et al.*'s model [1] is constructed according to the blue-print of Treisman *et al.*and Koch *et al.*model, it inherits both pros and cons of the previous technique. For example, knowledge about which the underlying features are predictive of fixation patterns is lost if responses from the features are integrated into a single topographical saliency map. With this featureless representation, it is no longer possible to measure the individual contribution of orientation versus colour and spatial frequencies. In other hands, it also has some significant gains; for example, pooling various features to derive a representation of saliency leads to greater consistency in predicting fixation for more types of images. This observation is intuitive since we have referred in the beginning of this section that a single feature might correlate well with eye-fixation data for only a subset of images. Uniting more feature maps in a featureless representation might as well push a chance of producing universal saliency maps which are consis-

tent with eye-fixation data of any input images. A distinct difference between Itti *et al.*'s model [1] and its predecessors is its concern about covert and overt attentions. In their descriptions, Treisman and Galade [24], as well as Koch and Ullman [35], mention how covert attention is achieved and evidences of overt attention are derived from foveal displays with controlled fixation. However, in the later work of Itti *et al.*[1], experiments are carried out in concern of both covert and over attention simultaneously without any effort of discriminating them.

Despite popularity of the FIT-like model such as ITT model [1], significant criticisms is levied against its definition of saliency. For example, its structures and computational methods is hand-crafted according to results of primate neurophysiology. The neurophysiology experiments show that V1 visual cortex tunes to different orientations, color opponency and intensity processed by centre-surround operators at various spatial scales. It inspires inclusion of equivalent components of the saliency map paradigm. Since the proposed model tries to resemble cells of visual cortex through investigating into their behaviours, they offer little explanation *why* a cell should have this structure. Moreover, it is significantly difficult to transfer these computational tasks into modulations of signals at a single cortex. Nevertheless, one possible connection between features typically found in the early cortical processing and quantitative principles is posited by Koenderink and van Doorn [60]. They suggest that early visual circuity is capable of sensing the visual patterns with respect to a principled definition. It indirectly confirms possible derivation of V1-style features usually employed in early attention models from receptive field in a principled geometrical framework. Noted that, the relationship between features employed in ITT model [1, 36] and geometric motivation of Koenderink *et al.*[60] only exists in abstract level.

The impetus behind the properties of features in early visual processing is partially supported by series of Linsker's work [61–64] on infomax principles. Linsker studied learning process of mapping input units to a set of outputs such that it maximizes Shannon entropy between inputs and outputs. When Hebbian style learning on a neural network is utilized, responses at units have properties reminiscent of those observed in early cortical processing. Moreover, those behaviours also replicate centre-surround operation configurations [61], specific combinations of angular and radial frequency [62], and cortical column-like organization [63]. A resemblance has been found between the learned receptive field profiles and those early visual features employed as building blocks of several early visual attention, but a connection has not yet been made explicit.

In the advancement of research in computational visual attention, the model of Itti *et*

al.[1] is remarked as the first complete implementation with sophisticated functions of hypothesized circuitry involved in attentional selection. Perhaps, the second most important contribution of the model is addition of various other features such as flicker, stereo disparity, shape from shading [36], gist [65] despite limited supports from neurobiological evidences. Therefore, it limits accountability of these features with respect to what constitutes saliency visual contents. Additional limitations of Itti *et al.*'s model are ignoring the issued of modulation throughout cortex and the accumulation of features into a featureless representation prevent determination of activating stimuli and its according features.

2.4.2 Frequency-based Saliency Method

Though Itti saliency model [1] is considered as a sophisticated method, it does suffer from serious disadvantages: high computational cost and variable parameters. Therefore, saliency computation is still open challenge and other methods such as Spectral Residual (SR) [52], phase of fourier transform (PFT), and An Information-Maximization method (AIM) [18] have been proposed.

In this section, an overview about saliency methods utilising frequency descriptors such as spectral residual (SR), phase of Fourier transform [2], etc is provided. The SR approach, based on the Fourier Transform, was proposed by Hou [52] with main contributions of being free of parameters and able to rapidly detecting objects. However, the framework is not easily extensible to include motion features and nearly impossible to calculate saliency maps for spatial-temporal events. Phase Fourier Transform (PFT) [2] and Phase Quaternion Fourier Transform (PQFT) [2] are developed to improve performance of spectral residual approach (SR) and include motion features in saliency calculation.

Phase Fourier Transform works on following principles, which can be easily understood by its 1-D example. Assumed that there are one pulse signal and one periodic signal, illustrated in three left plots of Figure 2.2, applying PFT will have different results, demonstrated by three right plots of the figure 2.2. In the case of pulse signals, significant spikes will be seen at the reconstructed signal, two top-right plots of the figure 2.2; meanwhile, there is no spike in a reconstruction of the periodic signal, a bottom-irght plot of the figure 2.2. According to the above demonstration, it seems that disruption of periodic signal is assumed to create saliency values. Analogously, similar principle can be deployed for two-dimensional image-like signals; for example, he PFT

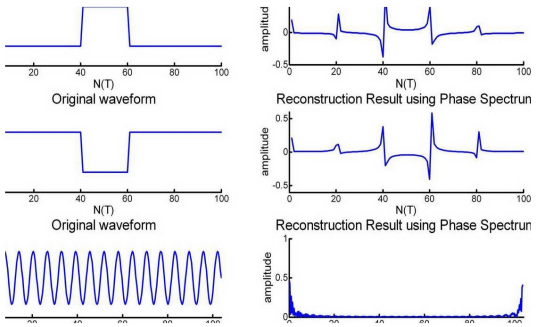


Figure 2.2: PFT method demonstration on 1-D pulse and periodic signals [2]

saliency approach of an 2-D image $I(x, y)$ can be mathematically described as follows.

$$f(x, y) = F(I(x, y)) \tag{2.4.9}$$

$$p(x, y) = P(f(x, y)) \tag{2.4.10}$$

$$M(x, y) = g(x, y) \mid F^{-1}[e^{i.p(x,y)}] \mid^2 \tag{2.4.11}$$

Where F and F^{-1} denote the Fourier Transform and Inverse Fourier Transform, respectively. $P(f(x, y))$ represents the phase spectrum of the image. $g(x, y)$ is a 2D Gaussian filter. In PFT, only intensity value of an input image $I(x, y)$ is used as the input signal. Besides intensity, there are many other features of images which have not been utilized in PFT. That motivates development of PQFT saliency method which can use up-to-four features at a time. PQFT is similar to PFT in core principles but has one distinctive feature: all calculations are done in Quaternion algebra. Input data are a combination of four different features: intensity (I) , red-green colour (RG), blue-yellow colour (BY), and motion (M) as follows.

$$I(t) = (r(t) + g(t) + b(t))/3 \tag{2.4.12}$$

$$RG(t) = R(t) - G(t) \tag{2.4.13}$$

$$BY(t) = B(t) - Y(T) \tag{2.4.14}$$

$$M(t) = \mid I(t) - I(t - \tau) \mid \tag{2.4.15}$$

$$q(t) = M(t) + RG(t)i + BY(t)j + I(t)k \tag{2.4.16}$$

After formulating the input signal or quaternion images, the Quaternion Fourier Transform (QFT) [66] is applied.

$$Q(t) = F(q(t)) \tag{2.4.17}$$

and Q(t) can be represented under its polar for form

$$Q(t) = \| Q(t) \| e^{\mu\phi(t)} \tag{2.4.18}$$

where $\phi(t)$ is the phase spectrum of $Q(t)$ and μ is a unit of pure quaternion. After setting $\| Q(t) \| = 1$, $Q`(t)$ only contains phase of frequency domain. The phase spectrum can be used to construct output saliency map by using inverse Fourier transform of Quaternion number.

$$q'(t) = F^{-1}(Q`(t)) \tag{2.4.19}$$

Later, 2-D Gaussian filter will be applied to smooth output saliency maps. The spatial-temporal saliency method using PQFT considers and process four features such as motion, colour, intensity and orientation in parallel. Thus, it is more economic in terms of computational load than ITT models. Moreover, it is independent of variable parameters like PFT and SR.

Guo *et al.*[2] use quaternion numbers as a holistic representation for early visual features such as colours, intensity, and motions. However, the quaternion framework can be extended for other types of data as well even though some are not neurobiologically plausible. That means usage of other image features which often appear in early visual processing. For instance, Schauerte *et al.*[67] proposes similar quaternion-based spectral saliency with weighted quaternion colour space components instead. Mathematically, such features can be written in PQFT mathematical framework as follows.

$$\mathbf{I_Q} = w_4 * \mathbf{I_4} + w_1 * \mathbf{I_1} + w_2 * \mathbf{I_2} + w_3 * \mathbf{I_3} \tag{2.4.20}$$

Since colour-map systems usually have three channels such as Red(R)-Green(G)-Blue(B), Luma(Y)-two chrominance-(U-V), the real component of quaternion number is usually redundant and eliminated. It results in pure quaternion state as shown in the Equation 2.4.21

$$\mathbf{I_Q} = w_1 * \mathbf{I_1} + w_2 * \mathbf{I_2} + w_3 * \mathbf{I_3} \tag{2.4.21}$$

According to weighted colour features, Schauerte *et al.*[52] reuse concepts of Spectral Residual or Spectral Whitening approaches to formulate their saliency methods, de-

scribed in the below equations 2.4.22 and 2.4.23.

$$S_{ESR}(I_Q) = \mathbf{h}_\sigma * \left| F_Q^{-L}[e^{\mathbf{R}_Q + \mathbf{E}_\gamma \circ \mathbf{E}_\Theta}] \right| \tag{2.4.22}$$

$$S_{ESW}(I_Q) = \mathbf{h}_\sigma * \left| F_Q^{-L}[e^{\mathbf{E}_\gamma \circ \mathbf{E}_\Theta}] \right| \tag{2.4.23}$$

where R_Q, E_γ, E_Θ are consecutively spectral residual, eigenaxis and eigenangles of the quaternion color feature while $\mathbf{h}_{\sigma_S}$ represents Gaussian filter with standard deviation σ_S and $\circ$ is a Hadamard operator. Moreover, Schauerte's model [67] analyses images in multiple resolution, which has not been considered in SR, PFT, PQFT. It is done by combination of the spectral saliency of the image at different scales. Let I_Q^m denote the quaternion image at scale $m \in M$, then

$$S^M = \mathbf{h}_{\sigma_M} * \sum_{m \in M} \phi_R(S(I_Q^m)) \tag{2.4.24}$$

Along with the proposal of weighted colour features for pure Quaternion frameworks, Schauerte *et al.*also evaluate its performance against other well-known saliency methods. More details about its mechanisms and experiments can be easily found in their publication [67].

2.4.3 Information-based Saliency Methods

Previous sub-sections have briefly looked into temporal and spectral representation of data in different saliency methods. Both types of data are directly collected as matrix of pixels from what may be perceived by human vision systems. In other words, the pixel values are analogous to immediate responses of visual cells when light beams reach retina. Conveniently, pixel values are directly employed as input signals; however, there might be additional pre-processing steps, capable of transforming data into more meaningful and useful representations. One of such ambiguous but feasible data types for visual saliency bases on definitions of information theory and probability. A proper probabilistic representation of visual context benefits computation of visual saliency since this approach inherits many possible frameworks for inclusion of prior knowledge as well as other contextual information. Moreover, several prior efforts [68],[69],[70] have argued importance of contexts in both "where" and "what" path ways for predicting salient locations as well as detecting or recognizing objects. Nevertheless, there are several ways in which contexts and prior information can be exploited towards generation of better computational saliency methods.

A brief overview of existing principles in probability-based saliency methods are pre-

sented here for better overall understanding of the research topic before they are described with greater details in the next chapter. In order to understand the probabilistic aspect of salience, the notion of context and variety of related issues needs to be considered since determination of saliency is specifically and heavily impacted by the context-generated likelihoods. According to Bruce and Kornprobst 's works [71], prior contextual information are defined by two elements: locations and properties of features. The rest of this section provides discussion on these two fundamental issues rather specific strategies for saliency computation since solid understanding about the elements will be important in consideration of any saliency methods based on probabilistic formulation.

Firstly, prominent influence of location information in a Bayesian probabilistic formulation is analysed. In several studies [72],[73], information about an object location on the global receptive fields provide strong prior for determining its interests. As there are evidences of significant correlation between eye fixation data and central regions [50], prior knowledge of object locations provide far more powerful clues for determining salience than other basic features or bottom-up measures. However, overall determination of saliency based on prior location have been recently found to be merely an artifact of gathering photographic stimuli and experimental conduction [50]. That said, in Schumann's study [74], the problems are due to normal composition and screening of photographic images, attentive subjects tending to be in the middle, which does not reflect conditions of free visual sampling. Since the challenge at hand is constructing an accurate model of human judgement for saliency in general scenes, not only artistically composed images, the strategy of using central bias as independent factor for saliency computation is unsuitable to be employed. Moreover, central locations of an object in the receptive field does not mean salient in free-sampling general cases [50]. Above evidences defy treatment of locations as independent factors and usefulness of location information in saliency determination. In contrast, joint likelihood of features and locations are considered as vital clues for solving the question since some events are standing out due to unexpected appearance of certain features at specific locations. For example, surveillance applications have strong prior assumptions about emptiness in a restricted or fenced areas but crowdedness in an open public area. If there is a sudden surge of movements around the restricted areas or disappearance of crowds in the public areas, the situation will catch the attention eventually. With this sort of argument, detection of salient events with respect to its location will be also useful in wide range of applications such as finding illegal turns by vehicles, left behind items, unexpected events, etc. In brief, joint modelling of features and locations will provide interesting and promising solutions for the problem of visual attention.

Secondly, properties of the environments where saliency computation takes places are discussed due to its richness of prior contextual information. Scales of environments vary from local surround regions [75],[4],[76] to the entire image [4],[77], to the manifold of natural images [72],[77]. If environments are considered as local surround regions, we will have probabilistic centre-surround operators, which are main contributions of this thesis. Moreover, local regions are also considered in multiple scales which includes coarse-to-fine features of local environments. Such treatments of environment have undoubtedly improved the accuracy of saliency estimation. Prior information of environments at levels between current scenes and manifolds of natural images is also promising to generate better judgement of saliency but this information is not considered in the current studies. For example, if walking in preserved tropical forests, people have prior expectations about seeing endless routes of huge trees and being surrounded by wild animals, mosquitoes, etc. They would be surprised if urban scenes such as high-ways, modern buildings are suddenly found in such places. Above scales of environments are chosen such that they are helpful for conducting experiments on database of natural images. Prior expectation itself need not be categorical or bounded by such artificial scales. It might also take on a more loose definition such as a prior based on recent experiment, past knowledges, or task definitions. Moreover, a rich contextual model might be configured by several likelihoods of various different levels of contextual abstraction which all contributes to the final assessment of visual salience.

As previous discussion mentions usefulness of information theory concepts like location-feature contexts, prior knowledges for providing clues about salient regions, it would be interesting to consider visual attention and saliency in informative framework. Attneave [3] has suggested this possibility but fell short of explaining why the framework may provide appropriate contexts for considering visual saliency problems. Furthermore, discussion about formulation of previous approaches is carried out later with more intuitive interpretation of visual attention or salience from an information theoretic point of view. Attneave's experiment suggests one aspect of information-theoretic ideals which seem to anecdotally agree with human intuition about saliency. According to her experiment, saliency of a region is propositional to likelihood of errors (or number of guesses in the hypothetical game) before observers approximately get description of the region on the basis of its context. In other words, saliency in information-based framework may be its Shannon surprisal or self-information. However, it is just one line of reasoning; other ways of defining saliency in information-related context are also mentioned thereafter.

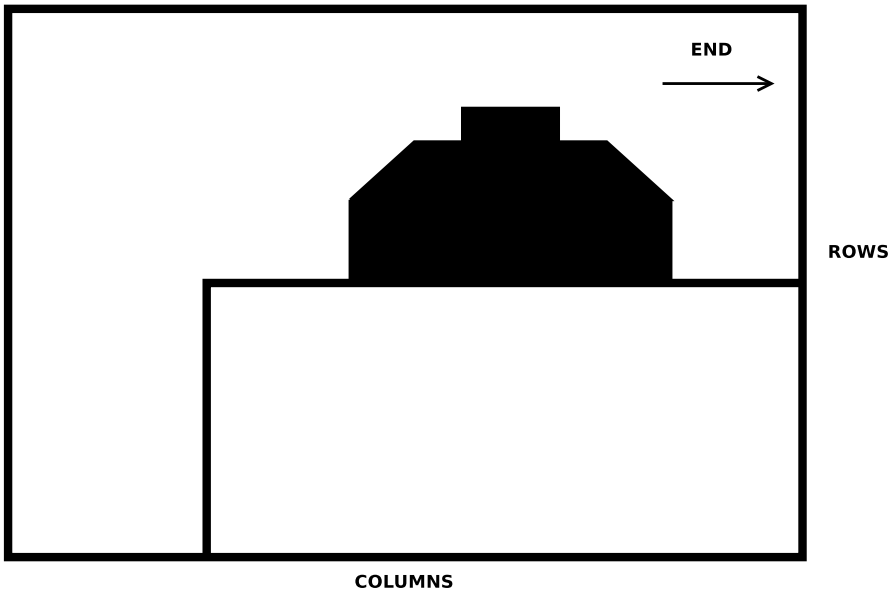


Figure 2.3: A abstract picture of an ink bottle on a desk corner [3]

Information-based Saliency of Attneave, 1954

In 1954, Attneave [3] had suggested that information aspects such as efficient coding and theory pertains to visual perception. Perhaps the relationship between information theory with visual attention and its intuitive role behind saliency computation are best introduced by using the author's imaginary experiments with following description and the associated figure 2.3.

> Consider the very simple situation presented in Fig 2.3. With a modicum of effort, the reader may be able to see this as an ink bottle on the corner of a desk. Let us suppose that the background is a uniformly white wall, that the desk is a uniform brown, and that the bottle is complete black. The visual stimulation from these objects is highly redundant in the sense that portions of the field are highly predictable from other portions. In order to demonstrate this fact and its perceptual significance, a variant of the "guess game" technique with which Shannon [78] has been employed for studying the redundancy of printed English. The picture might be divided into arbitrarily small elements which are "transmitted" to a subject (S) in a cumulative sequence, having him guess at the colour of each successive element until he is correct, ... If the picture is divided into 5O rows and 80 columns, as indicated, our S will guess at each 4000 cells as many times as necessary to determine which of the three colours it has. If his error score is significantly less than chance [2/3 X 4000 + 1/2(2/3 X 4000) = 4000], it is evident that the picture is to some degree redundant. Actually, he may be

expected to guess his way through Fig.1 with only 15 or 20 errors.

In the same manner with Shannon experiments of guessing English words in information theory study, Attneave's "guessing game" intends to highlight significant redundancy in natural visual stimuli, and human "players" seems to have insight knowledge about this redundancy. Though possibility of biological internal model of redundancy was not further investigated in Attneave's work, her experiment sets fundamental building blocks to the topic of this thesis. Still in terms of Attneave's experiments, the most interesting areas of the scenes are places which subjects only guess right after the greatest number of trials and errors. To further clarify this so-called definition of interest areas, lets consider analogous example in modelling redundancy in the English language and revealing that less frequent words within a particular context tend to provide essential information. For instance, the sentence from Bryan Adams' song "When you love a woman and tell her that she is really wanted" is still understandable if it is shortened as "love woman tell she wanted". However, another truncated form of the same sentence such as "When you a and that is really" completely destroys the original meaning. Apparently comparing to the later set of seven words, the former five words appear much less frequently. In the same manner, visual patterns with less chances of occurrence may have higher chance of being informative in a Shannon sense [79]. The definition of "interest" area can be alternatively interpreted such as *the saliency of visual signals may be gained by a measure of their contended degree of surprise*. The idea of considering information theory as backbone of visual attention and saliency theory is strongly supported by neurobiological evidences from Palm's work [80]. Palm outlines association of surprise characteristic with temporal structure of spike trains; therefore, it is amenable to consider the surprise or information in various aspects of early visual processing in computational abstract and neural levels. Such findings motivate further discussion presented in this chapter.

Let's consider a generalization of hypothetical experiments described by Attneave where observers need guessing contents of regions with arbitrary properties. Though the experiments are not carried out in practice, it is intuitive that observers make guesses on the basis of their surrounding context. For example in the Figure 2.4a, there are three blacked out circular regions consequently labelled A, B, and C. An observer can have general impression about the overall content of the image, and surrounding context strengthens beliefs of what contents are hidden under labels. Without any prior knowledges, most would guess white-cloud textures to be presented under these labels. As evident from the original image shown in the figure 2.4b, it can be seen how close the guesses come under labels A, B but how far the guess goes under the label



(a) Test Image **(b)** Test Image

Figure 2.4: Evaluation Test 2 - Moving patterns

C. If A, B are predicted easily, predictions of hidden contents under C need more effort due to its significant difference from expectation. Therefore, this region is also the most informative in Shannon sense. The imaginary experiments solidifies the proposal salient regions correspond to surprising and fundamentally informative in Shannon senses. In the latter part of this chapter, the computation of saliency values according to its informative content is formally described and examples are demonstrated.

A Local Entropy Saliency Measure

The notion of entropy is exclusively utilized in many visual attention works pertaining to information-theoretic considerations. Among different entropies, Shannon entropy is commonly used and the measure is defined as follows.

Given a discrete random variable X with possible outcomes $x_1, x_2, ..., x_n$, the Shannon entropy $H(x)$ is formulated as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log(p(x_i)) \tag{2.4.25}$$

And its continuous analog is

$$H(f) = -\int_{-\infty}^{\infty} f(x) \log(f(x)) dx \tag{2.4.26}$$

with f a probability density function on $\mathbb{R}$.

The above definition can be interpreted in many equivalent ways. According to physical work, entropy is related to a measure of uncertainty. From the above definition and interpretation, entropy is obviously maximized in the total uncertain system with

a uniform distribution. It is minimized in the total certain system with a distribution of random variable with only one possible outcome. Besides two extreme cases, other cases may simply be illustrated by the example of coin toss. For a fair coin, two possible outcomes (head/tail) are equivalent in terms of possible outcome (50%/50%); therefore, its entropy is maximized $-0.5 * log_2(0.5) - 0.5 * log_2(0.5) = 1$. For a biased coin with 90% of chance for showing up heads and 10% of chance for tails, it is less uncertain in guessing heads or tails; therefore, the entropy measure of uncertainty is reduced $H(X) = -0.9 * log_2(0.9) - 0.1 * log_2(0.1) = 0.469$. In other words, to a certain extent, the outcome of the coin-toss in later situation is less uncertain than the former one or entropy is reduced. Shannon entropy (H) possesses three important attributes: (i). H is maximized for a uniform distribution. (ii). H is a continuous function of the given probability distribution (iii). H is unchanged regardless of any types of regrouping given data.

Entropy can be interpreted in a slightly different way relating to an idea of surprise or self-information, $-log(p(x_i))$. In this interpretation, entropy is the average measure of surprise over all possible outcomes. For example, in case of the biased coin, it would be unsurprising if the coin came up heads. Therefore, the average surprise of biased coin is significantly lower than that of a fair coin.

Apart from the interpretation of entropy in the context of mathematical meaning, the measurement pertains to neural coding as well. An encrypting process refers to activity of relating some concepts or representation by others in order to gain better and shorter representation. Let's consider a hypothetical language with only four basic characters (a,b,c,d) instead of 26 English characters. Provided that there existed long documents written in this language, probability distribution of these characters would be estimatede reliably. For example p(a) = 0.5, p(b) = 0.25, p(c) = 0.125, p(d) = 0.12 are assumed to be probability values corresponding to the four basic characters. Among elementary coding schemes, binary encryption is widely utilized in any signal processing systems. Let's replace the alphabetical characters (a,b,c,d) by (00,01,10,11); obviously two binary digits are required for encoding each character. However, there may exist another coding scheme for even shorter representation. Now if one examines another possible encoding scheme of replacing (a,b,c,d) by (0,10,110,111). Two bits are no longer necessary for encoding each different number of bits are specially assigned for an correspondent character. For instance, only 1 bit is used for conveying letter (a), 2 bitcharacter, but s are needed for (b), or even 3 bits are required for (c,d). Then, average number of bits for conveying each character in this long context is given by 0.5 * 1 + 0.25 * 2 + 0.125 * 3 + 0.125 * c = 1.75 bits. In other words, the later scheme has provided better and more

compact code for the language in this case. Moreover, it reveals that entropy measures the average number of bits that is necessary to represent each random character, event or observation. It later becomes one of key points when entropy is put in the context of neural coding which would always seek optimal signal representation. This brief discussion seeks to highlight the meanings of entropy in encoding, but mathematical descriptions of the process in details are out of scope, as the necessary backgrounds are easily found in Shannon' classic publication in information theory [79].

In the context of visual stimuli, the entropy of image content is analogous to its relation in encoding except that a normal dictionary of characters is replaced by a collection of intensity values for image pixels within a local image patch. With given data from local neighbourhood, it is easy to derive distribution of intensity values and the corresponding entropy. For the sake of simplicity, a binary image is considered due. Even with only two "characters", there are three possible types of intensity distribution in arbitrary local regions: a patch of mixed black and white, a patch of predominantly white pixels, or a patch dominated by black ones. In the case of an equal number of black and white pixels, data of the image patch are distributed uniformly; therefore, its entropy is maximized. However, an image patch with one dominant type of pixels has much lower entropy because of less average bits necessary for encoding each pixel. Similar arguments are done for the case of gray-scale except that the number of bins are 255 types of pixels instead of 2 types. It can be noted that the ordering of pixel elements does not matter or change the estimated entropy in any case since only statistical distribution of values matters.

There has been considerable interest in information theory concerning visual attention and eye movement fixations. For example, Najemniik and Geisler [81] propose an ideal Bayesian observer for predicting eye movement such that the gain in information about a search target is maximized. The most significant contribution of this model is showing a strong relation between neurobiological visual search task and maximizing target information by shifting fixations locations. In the model, searching target is done by maximizing posterior probability of potential target locations given a random background that has the spectral characteristics of natural scenes. Correlation between model results and ground-truth data shows potentials of employing Bayesian machines as principle mechanisms of driving human eye movement. Furthermore, Najemniik *et al.*'s model also mimics inhibition-of-return process, found in human attention, which prevent two fixation points occupying the same location. It is implemented by embedding some coarse memory of past fixation locations and reducing likelihood of predicting those points as where the next target should be. Therefore, it limits a chance of

revisiting any pass fixation locations and large regions around them.

Lee and Yu [82] argue roles of saccadic eye movements in information maximization is to overcome constraints of retinal images in terms of region coverage. In other words, eye fixation deployment guides the coverage of retinal images such that it conveys the maximum of information. Though local entropy is still utilized, it is not estimated from simple features but a local hypercolumn representation of Gabor filters. The probability distribution is computed from response values of local hyper-columns and its surrounds with appropriate normalization; then, local entropy is easily figured out from Shannon's formula. Hyper-columns of Gabor filters are preferred to simple pixels due to observations that local contrast features from many transformations tends to correlate with eye-fixation patterns better than any low-level features. The local entropy measurement can be extended to conditional entropy in order to include a prior memory in the form of a mental mosaic. Obviously, prior memory is a strong factor in choices of saccades by discounting mutual information between any given target location and the existing mosaic image from entropy of that location. This model connects several psychological phenomena such as pop-out, inhibition of return, to longer visual experience and short-term working memory, but still utilize only basic local entropy or conditional one; however, its performance was never tested on eye tracking data.

Renninger *et al.*[83] establishes that Lee and Yu model as a plausible description of primate saliency computation by setting up experiments with human observers. In this experiment, the observers are shown two sequential sihouette and asked whether they are matched each other. When the experiment is conducted, eye movements of human subject are tracked to observe the strategy underlying fixation selection. The experimental results confirm a particular deployment of eye fixation so that local uncertainty is minimized. It is equivalent to maximization of information assumed that the "information" equals to local entropy. As mentioned several times before, maximization of entropy happens in case of a uniform distributions. Since probability distribution in Renninger's model inherits from Lee and Ye' one, they are responses from multiple Gabor filters. In this manner, highest information typically corresponds to regions of the shape silhouette with several edgelets in various orientations because those places give the most uniform distribution of multiple Gabor filters' responses. Similar to Najemnik and Geisler's work [81], it is confirmed that little benefit is gained for integrating information across successive eye fixations. At a single level, the strategy of maximizing information seems to be involved in an adaptive gain control mechanism or rescaling of neuronal input/output relation [84]. Though there is no direct relation between adaptive gain control [84] and Renninger's model [83], it at least encourages the informative

approach and the existence of an information maximization process in sampling during visual search. Furthermore, recent studies [85] have spotted a quantitative increase in information transmission of a cell from outside classical receptive field.

In addition to Lee and Yu or Renninger models, local entropy pertaining visual attention can as well be found in Kadir and Brady's work [22]. These authors calculate the local Shannon entropy by histogram of intensity or colour values in an arbitrary window. For each different window size, local Shannon entropy values are estimated from the corresponding local neighbourhood; therefore, we have an entropy versus scale curve at each location. For each curve, an effective entropy and scale are selected such that information is maximized. Besides local entropy, changes of probability density functions from two successive scales are measured such that the most significant difference is found. Finally, saliency values are given by product of local entropy and that maximum distance between probability density functions. Kadir and Brady typically visualize results by variety of circular regions at several spatial scales. In some regards, this model is unique in the way that it studies the scale variance of interested features across spatial scales.

Currently, local entropy is also utilized in the selection of features as precursor to object recognition by Fritz *et al.*[86]. Their models extract local features with entropy measurement, and Principal Component Analysis (PCA) is employed as a precursor to reduce dimensionality of local data. Though this work does not have direction relation to saliency, it employs local entropy to select a subset of features from available visual content.

All above approaches mentioned in this section pertain visual saliency to a measure of information and they advocate usefulness of local entropy as the basis for selecting salient features or driving gaze points. In other words, the first eye fixation points should be guided to locations or regions whose contents are encoded by the greatest number of bits or which contains the most information. Although the interpretation between saliency and information content is reasonable, there still exist some issues and twists. Stimulus from local content activates according neural responses which is assumed to reach its peak with a uniform distribution with respect to particular features. It implies that the local entropy measurement heavily depends on what features are used for stimulus representation, and it consequently affects the choice of gaze points. Parkhust *et al.*[50]'s work emphasizes on links between visual selective attention and features derived from specific combinations of angular and radial frequency. Even if adaptive features are found, there are cases where local entropy fails to predict eye fixations. It is probably due to local attributes of the measure, causing failure in detection

of interesting regions in global context.

A Self-Information Saliency Measure

In the previous section, several saliency approaches, based on local entropy measurement, were briefly covered, in addition to philosophical considerations that were used as rationales of local measurement. The interpretation of local entropy measurement in the visual attention context lies essentially in the hypothesis that a local region with the greatest number of bits of information certainly is the most attentive area. However, local entropy is not the only information measurement suggested for visual saliency, there exist other alternative methods as well.For example, Bruce *et al.*[18] introduces a slightly different view of the attention problem in an information theoretic context and contrast. It is inspired by observations of phenomenons in which notion of uncertainty provides misleading guide for visual attention; for instance, complete Gaussian noise results in highest local entropy while it says little for saliency and attention. In order to solve the above problems, Bruce suggests utilization of self-information technique to evaluate saliency value instead of local entropy. In order to know which measurement quantities should be chosen and why it should be selected, the attention phenomenon would be observed as follows: when facing input covering the entire visual field, it is necessary to decide what portions of the visual fields and which orders of these parts would be fixated. From this point of view, attention seems to be very reminiscent of elemental information-theory problem: *Given a limited capacity, what and how would messages be sent to optimize performance of information transmission ?* In this manner, visual attention can be regarded as biological transmission and communication systems. The overall visual field can be considered as message space including many information packets, attention strategy would choose and send appropriate packets, messages, for further analysis. Obviously, decisions of sending packets depend on the nature of the message space. It directly implies a basic problem of local entropy which does not possess any knowledge about message space. Therefore, the local entropy would mislead fixation of eye-deployments in the case of too small local neighbourhood. Details on cases where local measures fail spectacularly will be further discussed and elaborated in the rest of this section.

Let's consider a practical example wherein the entire visual field includes a white wall with an arbitrary black object in the centre. Obviously, sending messages pertaining to the black object would make more sense of the scene than sending those from the empty white wall. However, the situation would be varied if the background has complex texture patterns instead of a simple white and empty wall. It would make local

entropy of points from background larger than that of locations from foreground. That is when local entropy would mislead the information-based model of visual attention. Therefore, only local activity may no longer be sufficient to correctly measure interest in some certain context, and self-information is more necessary and better measurement for characterizing regions of interest.

The intuitive element of self-information is first ever mentioned in the work of Topper *et al.*[87] wherein basic features such as intensity, contrast, edges, etc. are considered as driving force of eye movements. One important conclusion achieved in his work is that remapping basic features to a Shannon's self-information measurement space produces greater correlation with eye movements. Though this information type is briefly defined and described in the beginning of this section, more detailed discussions are worthy due to its vitality in philosophical interpretation of information-based computational models, put forth in this dissertation. Though pioneering in utilization of Shannon's self-information measurement for saliency computation, Topper's work analyses incoming visual data by simple transformation; then, it leads to inaccuracy of estimation. Besides, it also suffers from methodological problems in gauging relation between deployment of eye fixations and saliency maps; then, overall performance of resultant saliency maps from Topper's work is relatively poor and inconsistent. However, basic ideas from his work are amendable to solving fundamental problem of local entropy in information-based selective attention models. In a specific aspect of visual saliency, Topper's work suggests self-information as transformation of neural modulated signals into visual saliency measure.

Visual attention can be modelled as mathematical framework, based on information theory; then, brain systems and activities are analogous to information systems. Let's consider that equivalence in a formal description of a general flowchart from information system. A data *source* needs transmitting to a destination through transmitters, receivers with additional arbitrary noise. Performance of an information system totally depends on the bottleneck in its channel, since the upper-bound of any communication system is limited with its least capable component. Obviously, it makes sending all information at once in parallel impossible, but rather pieces of source information are transmitted serially over time. The size of these information segments, once again, is governed by the upper-bound of the system. Given a limited time course, only a small portion of messages should be sent out from the dynamic information source.

Selective attention system suffers similar bottleneck problems of communication systems. With limited capacity of some channels, complex visuals can not be processed rigorously in an entire parallel manner. The problem is analogously overcome by se-

lecting subsets of visual content (message packets) from incoming scenes (information source comprising contents of visual fields) and relaying them to other areas of visual cortex for higher processing tasks. In agreement with standard communication systems, the first step of early visual attention may encode or transform imaging features into information space [44] which can be done by local entropy, self-information or many other possible ways. Then, messages are sent out in orders of decreasing information value to ensure maximum gain after each transmission. For example, a message contains a meaningful English sentence: *The University of Nottingham is expanding day by day*. Successful transmission of keywords like a university name *The University of Nottingham* and a verb expanding should be prioritized over common words like *is* or *day by day* because keywords are much more important than common words in conveying the sentence's idea. If messages are English separated words, message spaces are the English languages. In the context of this language, the appearing frequencies of common words *is* or *day by day* are much large than those of keywords *The University of Nottingham* and expanding. Therefore, a message or a word is more meaning full when it is rarer in a message space or a language. Whether a message space is involved in information quality assessment is the distinct separation between local entropy and self-information. While local entropy characterizes information amount of messages itself regardless of message spaces, self-information measurement does judge messages in the context of message spaces. In other words, self-information evaluates the contribution of information into overall knowledge after a message is sent out or a piece of visual content is selected and pushed up to higher cortical area. Let's denote X_n as possible events, messages, or visual content, its corresponding self-information is given by negative log of its probability $-log(p(X_n))$. In the biased coin situation wherein chances of heads and tails as outcomes are subsequently 90% and 10%, self-information highlights the fact that events of showing up tails is more informative or surprising. In communication system, self-information is a key measure for deciding which packets are transmitted. Analogously, the same measurement can help to choose which regions should be early attended and later processed for a given task.

Let's say Figure 2.5 appears on visual fields of observers' retina. There are groups of circles representing centre and surrounding environments C, S. Information-based saliency approach in general attempts to answer how much information in the central regions (C) contribute into its neighbourhoods (S). It is noted that circles are drawn for illustration, no specific shapes or groups of visual elements need defining in advance.



Figure 2.5

Self-Information vs Entropy

Basic definitions of local entropy and their relations to surprisal, self-information, in the previous parts of this chapter set platforms for comparing those information-based saliency measure. The differences between two concepts of information theory are highlighted in context of computer vision research for measuring regional saliency. In this manner, replacement of one measure by another leads to changes of a psychological visual attention phenomenon, known as the *popout* paradigm. Moreover, it involves different neural circuitry for either local entropy or self-information.

Let denote X as random variables from sample image features such as gray-scale, RGB

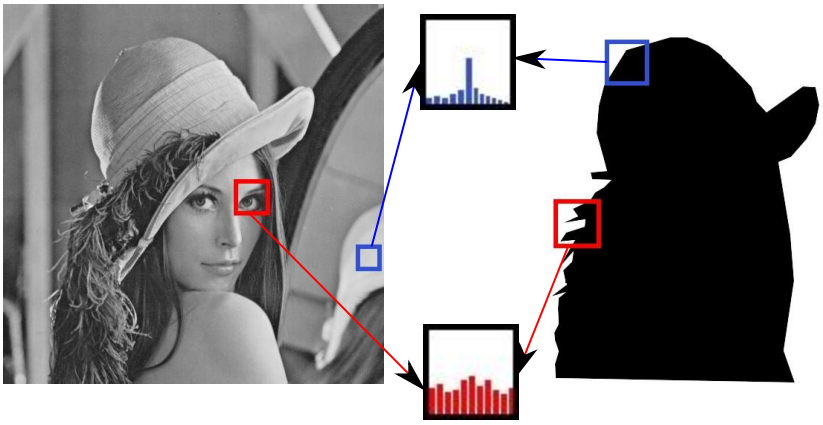


Figure 2.6: Left image with gray-level features, and **Right** image with orientation features are used for nearly uniform PDF

colors, etc. Practically, $X = [x1, x2, ..., x3]$ is high-dimensional features of an image patch and D is its corresponding probability density function of features inside the image patch X. The distribution can be observed by constructing histogram of intensity values or relative contribution of different orientations within local neighbours. Provided that histogram of features X is based on N equally-spaced bins, its entropy is given by "$-\sum_1^N D_i log(D_i)$" whereas D_i is frequency of features in the specific bin i. In this manner, local entropy characterizes how uniform the distribution D is with features from local patch X. The construction of distribution D greatly depends on which feature domains are chosen. When gray levels are chosen as representative domain, saliency measure obviously depends on the distribution of pixel values in considered areas. In the left image of the Figure 2.6, there are two regions of interests which are marked by red and blue rectangular windows. The red windows contains a fixation point and its neighbourhood meanwhile a blue window just covers a plain part from reflection of her hat on a mirror. In one hand, an eye's region contains several heterogeneous parts; therefore, its distribution of pixel values, histogram, is nearly uniform. In another hand, a plain region inside the blue rectangle is mostly homogeneous; then, its pixel values are only in a narrow range, and correspondent distribution is consequently highly non-uniform. Similar analysis can be done for orientation features of a silhouette image of Lena sample, at the right side of figure 2.6. A red window involves regions with complex structure and near uniform distribution of orientation; meanwhile, a blue region considers a simple shape with non-uniform distribution of orientation. In both sample images, regions under red windows has higher local entropy than an area under blue window. It is noted that silhouette images are employed in the study of Renninger *et al.*[83] with self-information as saliency measures, $-log(p(X))$. In their work, saliency value is measured by raw likelihood of the specific n-dimensional vector of RGB values. In this case, $p(X)$ is probability of the event X due to observations of n-dimensional feature vectors from a patch or a region around X. The probability $p(X)$ is the ratio between a number of patches with n-dimensional vectors over several patches. In addition, the negative log or self-information measure behaves like local contrasts whereas the entropy represents local activity. Distinction between two types of information-based measures is clearly shown in Figure 2.7. A major portion of the figure is covered by grass textures; however, there is an empty white space in its middle. Obviously, human eye fixation will prefer white markers to other outside green textures. If smaller red windows identify regions of interests (ROI), entropy of a red window 1 is smaller than that of a red window 2; meanwhile, self-information from features of a red window 1 is larger that of a red window 2. However, if big blue windows are utilized, local activity or entropy under a blue window 1 is bigger than

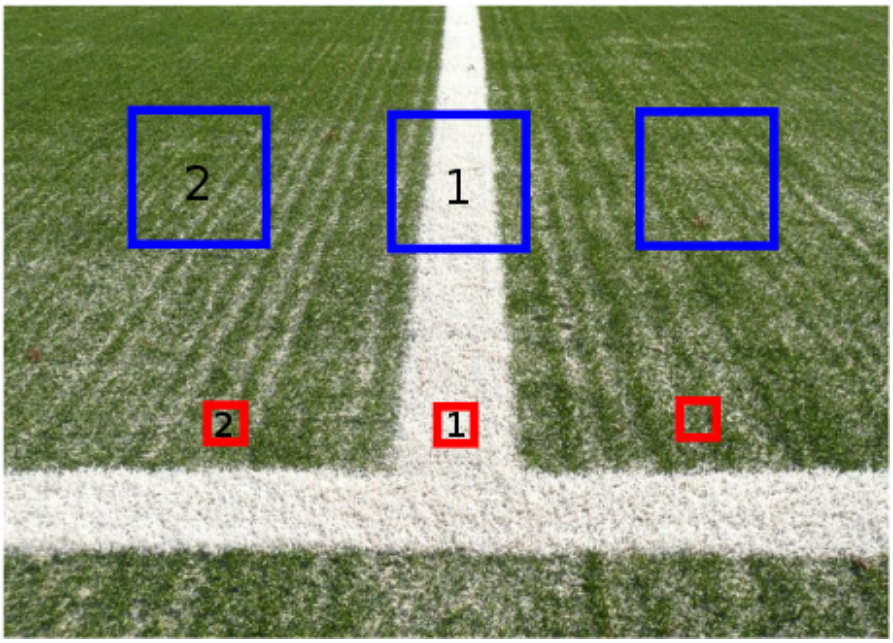


Figure 2.7: Example that highlight distinction between self-information and local entropy

that of a blue window 2 and so are the self-information. While self-information only accesses how informative a specific region is against the rest of the image, the local entropy is average of information over only this local region. Therefore, local entropy values heavily depends on correct choices of scales while the same issue does not affect self-information much. The issue of scale relates to receptive fields in a human visual attention process, and such cortex biologically tunes a region where entropy or self-information are computed on the basis of inter-neuron connections.

A Decision Theoretic Saliency Measure

In the previous section, histogram of image features are employed for estimation of local information. Both self-information and entropy (average of self-information over a region) do not involve contextual parameters or information; therefore, it lacks biological plausibility regarding to centre-surround contrasting behaviours of visual cortex. Considering such issues, Gao *et al.*[88] formulated a discriminant definition of saliency given knowledge of surrounding contexts. The salience is considered as the power of some set of features $\mathbf{X} = [x_1, \ldots, x_d]$ to distinguish between observations sampled from central and surrounding regions. At a location l, the feature set X is classified into either class $Y_l = 0$, corresponding to samples drawn from a surround region, or class $Y_l = 1$ corresponding to samples drawn from a smaller central region. The value of saliency is then formulated as mutual information between distributions of features X

and classes Y, $I(\mathbf{X}, Y) = \sum_{i=1}^{d} I(X_i, Y)$. Joint information estimation of two distributions are computational intractable, especially in high-dimensional feature spaces X; therefore, relaxation of independences between different dimensions is assumed. In addition, a few computation tricks are applied for computational feasible solutions. For example, conditional distribution of features X given class Y are assumed to fit a Generalized Gaussian distributions to simplify computational processes and features are first pre-processing by Gabor-like decomposition.

Discriminant theory does not blindly include contextual information to satisfy centre-surrounding behaviours of biological vision systems since this modelling configuration is also supported by neurophysiological researches. As widely known, primary visual cortex (V1) includes both simple and complex cells [89],[90]. Simple cells are classically considered as receptive fields at the neural levels; therefore, several studies focus on stimuli's incidents and treat those cells as cascades of a linear filter and non-linearity rectifier. However, important properties of simple cell behaviours such as gain control, were only recently found; hence, divisive normalisation of a cell response has to be added as an additional stage [91]. On the other hand, complex cells are usually modelled as integrating units for squared and half-rectified outputs of linear units of the same orientation. The optimal discriminant solutions of saliency is compatible with the combination of complex, divisive normalisation, and simple cells, called the *standard V1 architecture* [90].

Besides evidences from neurophysiological studies, the discriminant saliency model is found to be suitable explanation for psychophysical data available in literature [92]. For examples, the model performs consistently in searching asymmetries, a phenomenon well-documented by Treisman *et al.*[24]. That said, a unique feature of targets intrigues human attention; meanwhile, a target defined by lacks of features is difficult to be found. In other words, only a si ngle target feature, which are absent in background, can be easily found pre-attentively while, target-background differences due to a conjunction of features can not be resolved without top-down processing [24]. It is easily understandable if neural structures of discriminant framework are considered. A single feature of target, absent from surrounding environments, is easily distinguished by simple cells; meanwhile, involvement of a feature junction requires pooling, normalisation steps. Both qualitative and quantitative simulations using different patterns and Weber's law, have been conducted by Dashan Gao *et al.*[92] to confirm similarity between behaviours of their computational model and searching asymmetries. Generally, it is reasonable to conclude that discriminant theory is biologically plausible in terms of neural circuitry and consistent with psychophysical findings.

A Bayesian Surprise Saliency Measure

In visual attention research, definitions of contextual information can be relaxed to adopt many forms of mathematical modelling. In the prior section, it has been defined as features X_0 corresponding classes Y_0. However, it is just one and obviously not the only way of interpretation. In the surprise model [19], surrounding contexts are defined as prior knowledge M in a Bayesian framework; meanwhile, new data are denoted as D. Then the posterior probability distribution given a new data, $P(M|D)$ can be written as follows.

$$P(M|D) = \frac{P(D|M)}{P(D)} P(M) \tag{2.4.27}$$

According to this equation, D is the main driving force to change the prior distribution $P(M)$ to the posterior distribution $P(M|D)$. In other words, D is an operator channelling information from the distribution space to the model space. Thus, the transmission can be quantitatively measured by considering distance between the prior and posterior distributions. Different from Shannon's communication information which measures self-information amount of the system at once time, this quantity measures differences of information if new data are received. To distinguish two entities, Itti *et al.*[19] name their proposal as surprise information, or just surprise.

$$S(D, \mathcal{M}) = d[P(M), d(M|D)] \tag{2.4.28}$$

where d is a general operator for measuring differences of two distributions. As there are different ways of quantifying dissimilarities between probability distributions, the Kullback-Leibler divergence, a standard well-known theoretical measure, is chosen due to its invariance to reparameterizations.

$$\begin{aligned} S(D, \mathcal{M}) &= K(P(M), P(M|D)) && (2.4.29)\\ &= \int_{\mathcal{M}} P(M) log \frac{P(M)}{P(M|D)} dM && (2.4.30)\\ &= -H(P(M)) - \int_{\mathcal{M}} P(M) log P(M|D) dM && (2.4.31)\\ &= log P(D) - \int_{\mathcal{M}} P(M) log P(D|M) dM && (2.4.32) \end{aligned}$$

where H denotes the entropy. Alternatively, the surprise information can be defined average log-odd ratio of multiple single models.

$$S(D, \mathcal{M}) = \int_{\mathcal{M}} S(D, M) P(M) dM \tag{2.4.33}$$

where (2.4.34)

$$S(D, M) = log \frac{P(M)}{P(M|D)} \tag{2.4.35}$$

The average of single-model suprise is derived according to Bayesian rule with respect to the prior distribution over the model class. Surprise is also equivalent to the free energy of the negative log-posterior at a temperature $t = 1$ according to the prior distribution over the space of models [93] in terms of statistical mechanics terminology. As distinguishing from Shannon information, surprise information introduces its own unit - the "wow", the surprise corresponding to a two-fold variation between the piror and posterior distribution $-log_2 \frac{P(M)}{P(M|D)}$ in a single-model M. As measurement of dissimilarity, this log-odd ratio or "wow" unit can be positive or negative, corresponding to increment or decrement of observers' belief in a model M give a new dataset D. When simultaneously considering all models $\mathcal{M}$, we get averaged surprise or a total number of wows as an always positive relative entropy.

Previously we have discussed about formulation of bayesian surprise, its inclusion of surrounding contexts and differences between Shannon's information and Surprise information. The importance of such quantities in visual saliency computation is worthy for further discussion about their pros and cons. In accordance with a theory of outlier detection, surprise are defined as rare events with a low occurring probability. Such definition of surprise is closely related to Shannon's self-information ($I(D) = logP(D)$); however, there exist cases where Shannon's self-information and Bayesian surprise provides distinct results. The differences will be illustrated through two extreme examples corresponding to high Shannon's information and low surprise, and vice versa.

Many-bits-but-few-wows case: the extreme example usually happens when no inputs are unlikely for any single-model M. It is the case of randomly distributed data as input; then, Shannon's self-information becomes arbitrarily large or event infinitive

$$P(D) = \epsilon \qquad \ll 1 \tag{2.4.36}$$

$$\Rightarrow I(D) = -log(P(D)) \equiv \infty \tag{2.4.37}$$

Meanwhile, the surprise is always zero as $P(M_i|D) = P(M_i) = P(D) = 1/N$.

$$S(D, M) = log\frac{P(M_i)}{P(M_i|D)} = log(1) = 0 \tag{2.4.38}$$

$$\Rightarrow S(D, \mathcal{M}) = \int_{\mathcal{M}} S(D, M_i)P(M_i)dM_i = 0 \tag{2.4.39}$$

In this situation, D, a strong out outliner, is not necessarily valuable in terms for information for systems. In other words, D does not provide any useful information for discriminating between different models M_i. Therefore, random input D carries no surprise and it leaves the observer's expectation unchanged.

Few-bits-but-many-wows case: Lets take a simple example for illustrating such case.

$$P(D) = \frac{a(1-a)N}{N-1} \tag{2.4.40}$$

$$P(M_i) = \frac{1}{N} \tag{2.4.41}$$

$$\Rightarrow I(D) = \lim_{N\to\infty} P(D) = -log[a(a-1)] \tag{2.4.42}$$

$$\Rightarrow S(D) = \lim_{N\to\infty} alogNa + (1-a)log\frac{(1-a)N}{N-1} = alogN + H(a) \tag{2.4.43}$$

For large number of data N, Shannon's information of this case converges to the constant value; meanwhile the surprise converges to $alogN$. In other words, we achieve a arbitrarily large amount of surprise in combination with a finite Shannon's information. Thus data with such mathematical property are missed by outliner theory but well detected by Bayesian surprise due to significant change from prior to posterior distribution. Regarding this example, it is clear that whether the data is informative or not depends on more that just probability alone. At least, information of prior events need taking into consideration to make generated surprise index or saliency more robust and inconsistent across many events. That is exactly an advantage of surprise information over Shannon's information.

A Free Energy Saliency Measure

From the idea of modelling attention as information systems, prominent concepts in this line of research have been reviewed briefly. Regardless of self-information, average-information (entropy), or Bayesian information, it all depends on assumption about relations between rich information and human attention. However, the previous works are lacking one important element which is neuro-biological plausibility. In other word, they did not exposit how such their proposals operate at levels of biological circuits ex-

cept for Gao's discussion about relations between his mathematical proposal and the standard V1 scheme [92]. Nevertheless, his justification may be correct for the particular framework of discriminant saliency, it does not hold true for general information-based approach. To put the research line in cement, it is necessary to find neuro-biological evidences to support validity of algorithmic proposals.

Like most ground-breaking fundamental ideas in natural sciences, the theory of unifying informative saliency approaches and brain activities are somehow relating to physics and physicists. Friston, a profound neuroscientist and authority on brain imaging, has bravely initiated and tirelessly contributed to his free-energy principle and Bayesian brain. On the course of development, Friston regularly mentioned usefulness on free-energy principles in interpretation of attention, prediction, perception phenomenon [94],[95]. Moreover, through the principles are explained the operations of neural circuits for the above psychological concepts. In his writing for Frontier in Human Neuroscience journal [96], he has defined that attention is an optimal process of synaptic gain to represent the precision of sensory information during hierarchical inference. Furthermore, the neurobiology of attention is assumed to involve activity-dependent changes in synaptic gain wherein neuronal activities corresponds to external states while synaptic gain describes precision (prediction error). In general, an architecture of attentional processing can be simulated as Bayesian optimal inverse or recognition schemes with state-dependent noise.

Though mathematical principles of free-energy and Bayesian brain are the main interests, it is worth first mentioning about supporting neuro-biological evidences to solidify its biological plausibility. Attending to object often includes looking and locking it at the fovea (the central area of retina with highest density of visual cells). However, attention can be independently eye movement under the Posner's paradigm [97]. Though experiments involving human eye-movements, visual cues, human attention are suggested to locate salient regions even in absence of gaze shift. Two types of cues used in the Posner paradigm - central and peripheral- are directly relating to endogenous and exogenous cuing consecutively, which in turn fit well with bias competition theory. Exogenous cuing is thought as a bottom-up bias with prior expectation of an external event; while, endogenous cues are mediated by top-down bias in forms of directional instruction to locations of salient events. Furthermore, Decaix [98] and Risko *et al.*[99] demonstrated the performance and reaction time of observers in finding salient locations are not affected by knowledge of cue-target relationship as well as the proportion of valid cues over all trials. The experimental results implies that Poster paradigm might not involve any high-level cognitive processing. In other words, it

directly suggests pop-out-like behaviours of observers for cue-validity effects in Posner's experiments are fully accounted by low-level simulation of perceptual processing. This observation inspires development of Bayesian-optimal processes to explain such effects. As evidences point to Bayesian modelling of attention, Friston [100] develops such a system with free-energy formulation. Moreover, his proposal can be thought as a form of evidence accumulation or generalized predictive coding [101]. Free-energy is a bound on surprise, a log-evidence for the brain generative model of its perception. Surprise can be thought as uncertainty of the brain model on a particular state, and inversely there is precision measurement increasing with certain about the same state. That precision is mediated by the post-synaptic gain of sensory or prediction error-units, which means brain models perceive the outer world by continuously modulating gains of activity-dependent synaptic cells in forward connections. That is how Bayesian-optimal problem and predictive coding can directly model attention, a brain mechanism for visual activities.

The rest of this section focuses on reviewing basic mathematical principles of free-energy for attention proposed by Friston [9]. Given a brain model m and generalized sensory sates (concatenation of several states)$\tilde{s} = s^0 \oplus s^1 \oplus s^2 \oplus \ldots \in S$, this average uncertainty (entropy) of the model about these states is written as follows.

$$H(S|m) = -\int_p (\tilde{s}|m) \log p(\tilde{s}|m) d\tilde{s} \tag{2.4.44}$$

Under ergodic assumption, it is proportional to long-term average of surprise or negative log-evidence, $-\log p(\tilde{s}(t)|m)$.

$$H(S|m) \propto -\int_0^\infty \log p(\tilde{s}|m) dt \tag{2.4.45}$$

In order to minimise such entropy (uncertainty about the state of outer world perceived by the brain model m), the accumulated log-evidence is as well need suppressing. Though such sensory entropy can not be computed directly, an upper bound of that entropy, or free energy, can be created $S(\tilde{s}, q) \geq H(S)$. Along with this bound, a recognition density $q(t) := q(\mathcal{V})$ is introduced on the causes of sensory signals, compromising time-vary states $u(t) \subset \mathcal{V}$ and slow varying parameters $\phi(t) \subset \mathcal{V}$. The bound, a path integral of free-energy $\mathcal{F}(t)$, is created by simply adding the recognition

density to surprise.

$$\mathcal{S} = \int dt \mathcal{F}(t) \tag{2.4.46}$$

$$\mathcal{F}(t) = -\log p(\tilde{s}(t)|m) + D_{KL} \tag{2.4.47}$$

$$D_{KL} = \langle \log(q(\mathcal{V})) - \log(p(\mathcal{V}|\tilde{s}, m)) \rangle_q \tag{2.4.48}$$

The non-negative Kullback-Leibler divergence D_{KL} is zeros or equality state is reached when $q(\mathcal{V}) = p(\mathcal{V}|s,\tilde{m})$ is the true conditional density. It implies that strategy of minimizing free-energy by changing $q(\mathcal{V})$ makes the recognition density approximate conditional density on sensory causes, a Bayesian-optimal recognition. The free-energy can be evaluated easily in terms of recognition density and a generative model $\mathcal{L}(t)$ entailed by m.

$$\mathcal{F}(t) = \langle \mathcal{L}(t) \rangle_q - \mathcal{H}(t) \tag{2.4.49}$$

$$\mathcal{L}(t) = -\log p(\tilde{s}(t), \mathcal{V}|m) \tag{2.4.50}$$

$$\mathcal{H}(t) = \langle \log q(t) \rangle_q \tag{2.4.51}$$

The free-energy $\mathcal{F}(t)$ here is difference between energy $\mathcal{L}(t)$ expected under $q(t)$ and $\mathcal{H}(t)$, the negative entropy / negentropy of $q(t)$. The energy $\mathcal{L}(t)$ shows a surprise about sensory signals and causes under a generative model. Then, few computational tricks such as a Gaussian recognition density are used in order to derive explicit solution of recognition dynamics for expected sensory states and model parameters. Utilising the derived formula, free-energy as well as a bound on surprise (log evidence) are minimized, which corresponds to perceptual inference and learning respectively. The derived formula is not clearly stated in this section because its precise form depends on the energy $\mathcal{L}(t) = -\ln p(\tilde{s}, \mathcal{V}|m)$ which are in turn associated with a particular generative model.

2.5 SALIENCY APPLICATIONS

Computational models in the previous section are mostly designed to attract interests from natural images to demonstrate their simulation of human visual systems. However, these computational models might be a directly useful solution for many computer vision applications if they are slightly modified for specific problems rather than extraction of interesting points or regions. In order to utilise saliency methods as core solution rather than pre-processing steps, Feature Integration Theory (FIT) [1] and

Phase-based [2] models (PFT) are proposed as solutions for traffic sign detection and lane-mark extraction accordingly. As both FIT and PFT frameworks are previously described in Sub-Sections 2.4.1 and 2.4.2, Sub-Sections 2.5.1 and 2.5.2 focus on describing how these models are modified to solve practical problems without repetition of their details.

2.5.1 Feature Integration Theory for Colour-Based Traffic Sign Detection

Colours of traffic signs, as shown in Figure 2.8, have been well regulated (for example red means stop, yellow means danger) and even the paint covering sign boards corresponds to specific wavelengths[39]. As stated in a manual [102], red, orange, yellow, green, blue, violet, brown, and achromatic colours are widely used to paint traffic signs. However, violets and achromatic colours do not appear much on roads; therefore, only six discriminating colours are usually investigated. These colours can be utilised to tune Feature Integration Theory (FIT) saliency method as a solution for traffic-sign detection. Those colours are clues for generating "traffic-sign feature maps". Along with general colour, intensity, orientation maps in FIT presented in Sub-Section 2.4.1, the traffic-sign feature maps contribute and bias saliency maps to focus attention on regions with these signs.



(a) Red **(b)** Blue **(c)** Yellow **(d)** Green **(e)** Brown **(f)** Orange

Figure 2.8: 6 dominant colours used in traffic signs

Proposed Method In order to modify general FIT sliency architecture, described in Sub-Section 2.4.1, for traffic signal detection, the original FIT architecture needs to be enhanced with top-down bias. The modified FIT includes not only general features like colours, intensity, and orientation but also traffic-sign features such as 6 standardized colours of traffic signs: red, blue, yellow, green, brown and orange, as shown in Figure 2.9. Provided that red traffic sign is the wanted object, a conspicuity map of red colour would dominate the others when all conspicuity maps are merged into a final saliency maps. Moreover, the modified FIT model is quite flexible since the number of features can be controlled. For example, if only red traffic signs need to be found, only a red conspicuity map is involved in FIT. Therefore, all attention is focused to salient red traffic signs.

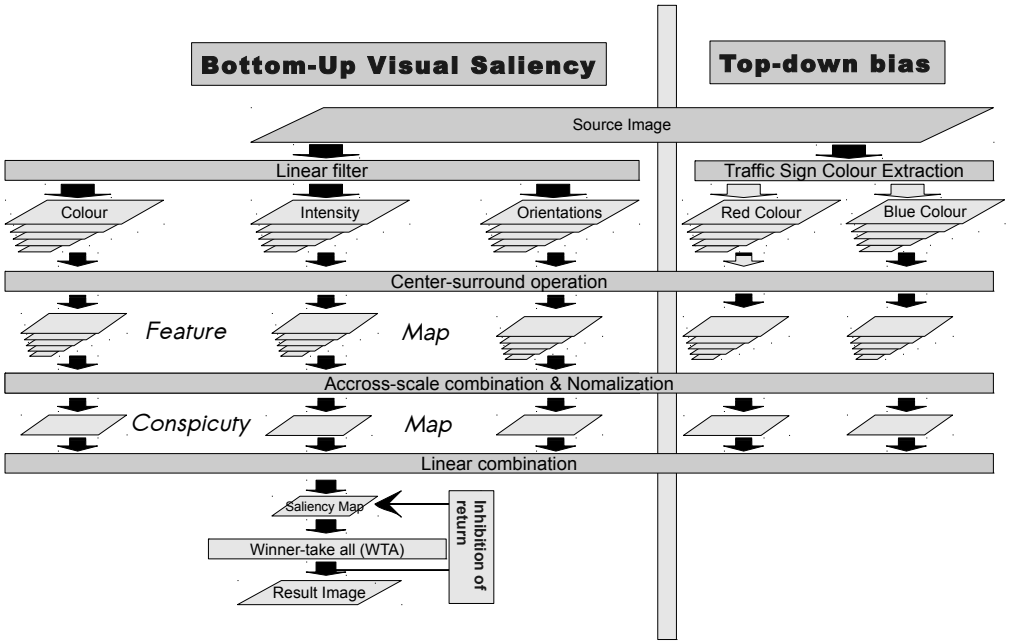


Figure 2.9: Modified architecture of a model:Saliency-based visual attention for Traffic Sign Detection

Standard Colour of Traffic Signs Colours of traffic signs always follow colour codes whereof each has its own meaning, defined in transportation law. Moreover, each colour is as well clearly defined in term of values in RGB colour system. For example, in the UK, the colours of traffic signs are regularized and published by Department of Transportation [102] as follows. There are 6 discriminating colours which are adopted internationally; therefore, they can be used as colour-based features of traffic signs . Table 2.1 specifies RGB values of these six basic colours.

Table 2.1: Main Colors of Traffic Sign in RGB[7]

Colors	R	G	B
Red	236	29	36
Blue	0	121	193
Green	0	112	60
Yellow	255	212	193
Brown	137	46	52
Orange	247	147	29

Since these colours in Table 2.1 are formalized, colour pixels from traffic-sign images, as shown in Figure 2.8, should possess similar RGB values to above standardized values. However, RGB values of same traffic sign colour might slightly vary across different images due to changing illumination, glares or other external factors. Therefore, RGB value of traffic-sign colour in captured images is an average value with an error rate.

If mean and variance (errors) of these colours are known, they can be used as guiding colour values for traffic sign object detection by bottom-up saliency methods. With these parameters of averaged traffic-sign colours, conspicuty maps, which favour those colours, are generated. A conspicuity map can be computed as HUE distance between colour value of a pixel and standard colours of traffic signs [8]. The HUE difference between a pixel and a standard traffic sign colour can be calculated as follows.

$$h(c) = \exp\left[-\frac{1}{2}(c-\mu)^T \sum^{-1}(c-\mu)\right] \tag{2.5.1}$$

Colour Extraction Besides external factors like illumination, colours are perceived differently by human eyes and computers. Through human eyes - a biological visual sensor, a certain quantization level of colours can be recognized, for example, red, green, blue, yellows, pinks, and other dozens of colour names. However, in accordance with computers, all colours are nothing more than numbers, or points in colour maps, and their accuracy of differentiating can be $\frac{1}{2^{24}}$ if 24-bit colour is used, or even $\frac{1}{2^{48}}$ if 48-bit colour is utilized. Ironically, computers have better presentation of colours than human beings does. In other words, it has more accurate definition or range of values for a specific colour.

Though RGB colour representation is the simplest but most well-known colour system so far, it does not necessarily mimic human colour perception perfectly. For a traffic sign detection task, Gao.et.al [39] shows that CIECAM (CIELab) does the best job among other colour systems. Therefore, CIELab is put to use for representing colour values of traffic signs, and all statistical measurement will be found in this colour space as well.

In order to get boundary values for traffic signs colours in Table 2.1, a database of different traffic signs has been collected and these images are classified into different categories according to their traffic sign colours. Under that classifying system, there are six different folders corresponding to six common colours appeared in traffic signs: red, yellow, orange, blue, green, and brown. Unique colour values will be extracted from these traffic signs to find out means and variance of their distribution.

The main objective of colour extraction steps is finding parameters of unique colours in traffic signs. Colour identification seems to be easy task for human beings, it is not simple task for machines. For example, when pixels of two very similar yellow colour are observed, they may look the same but actually have different values in colour spaces. Therefore, a yellow pixel of a traffic-sign does not necessarily has the same value stated in the Table 2.1, and there are some ranges of RGB values corresponding

to specific traffic sign colours. Since only chromatic data are of interest, colours can be well represented by two channels a and b of CIELab colour maps instead of RGB. Then, ranges of a and b values in accordance with main traffic sign colours can be learnt from traffic-signs images as training samples, as shown in Figures 2.8. Distribution of yellow colours, as shown in Figure 2.10, shows that chromatic data densely concentrate in a particular area of CIELab colour maps.

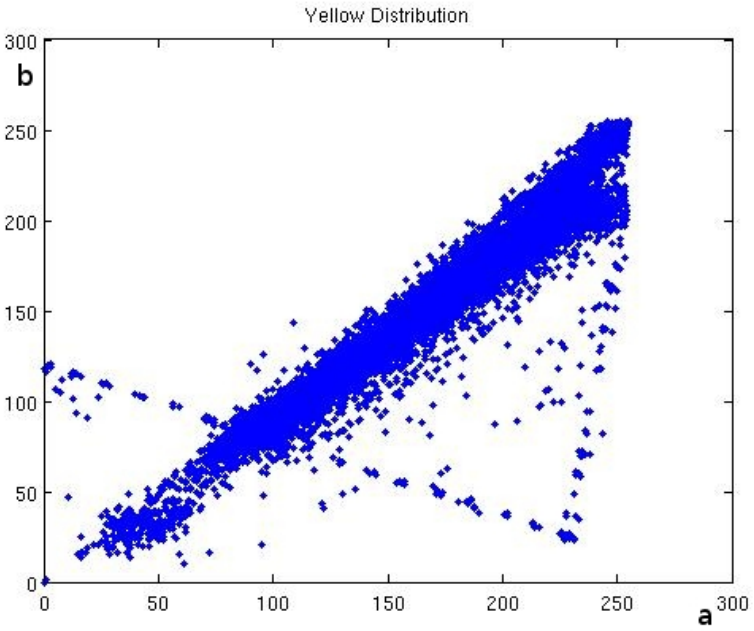


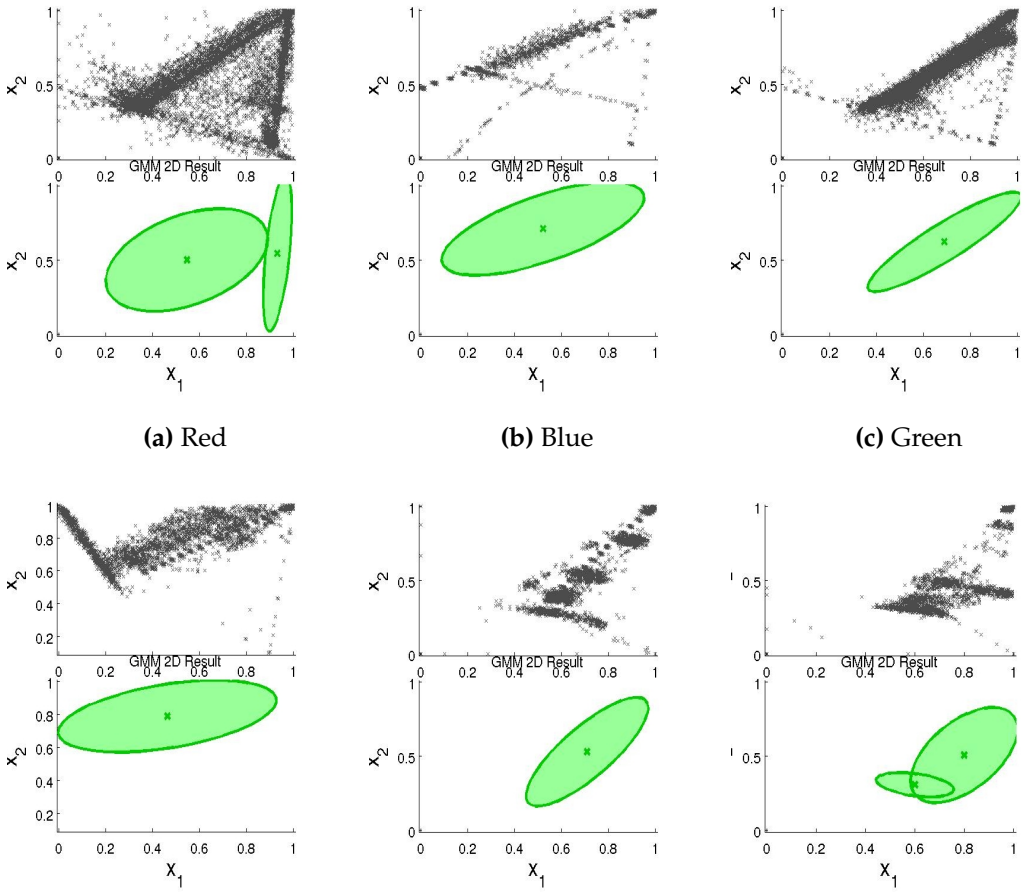
Figure 2.10: Yellow Colors Distribution in (a, b) space of CIELab

Colour Distribution A mathematical model must be applied on collections of colour values to learn their distribution parameters. Gaussian Mixture Model (GMM) [103] is assumed to be a model for these colour values and Estimation & Maximization Algorithm (EM) is utilized as an optimization method for corresponding distribution parameters. GMM [103] is mathematically described as follows.

$$p(x) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)\right] \tag{2.5.2}$$

GMM assumes that any distribution is a mixture of many Gaussian Distribution (GD) [103]. In this case, each GD represents one set of colour values which corresponds to a single standard colour of traffic signs, as shown in Table 2.1. If traffic-signs images are randomly chosen, values of traffic-sign colours are safely assumed to be randomly distributed. This randomness satisfies a deploying condition of GMM, which ensures

convergence to true parameters of the traffic-sign colour distributions. With GMM model, Estimation & Maximization (EM) step is utilized to compute means and variances for each distribution of a traffic sign colour. As traffic-sign colours have six basic colours, there are six Gaussian Distributions in this GMM model. Therefore, EM process is initialized with six different categories. Then, EM is iterated until all data are categorized into those six distributions optimally and corresponding parameters are identified. Six values of mean (μ_x, μ_y), variance (σ_x, σ_y), and correlation ρ values, as shown in Table 2.2, are statistical parameters of six traffic-sign colour distributions. μ_x and μ_y, α_x, and α_y are central points and radius along x and y axis of most significant eclipse, as shown in Figures 2.11a-2.11f.



(a) Red **(b)** Blue **(c)** Green

(d) Green **(e)** Brown **(f)** Orange

Figure 2.11: Traffic Sign Colour Distribution and Classification. (x_1, x_2) is another representation of (a,b)

.

Saliency Map With parameters of six traffic-sign colour distributions, six correspondent conspicuity maps can be generated according to Top-Down bias part of the mod-

Table 2.2: Major Traffic Sign Colour Classification Parameters

Colours	Red	Green	Blue	Yellow	Brown	Orange
ρ	2.36	1.31	0.79	2.12	0.98	1.70
μ_x	0.55	0.52	0.69	0.46	0.71	0.80
μ_y	0.50	0.71	0.62	0.79	0.53	0.51
σ_x	0.20	0.24	0.16	0.27	0.13	0.12
σ_y	0.20	0.17	0.16	0.12	0.20	0.18

ified framework as shown in the Figure 2.9. These maps contain Hue distances, computed by the Equation 2.5.1, between colour values of each image pixel and distributions of traffic sign colours. In other words, it represents likelihood of each image pixel belonging to a traffic-sign colour distribution. As c in the Equation 2.5.1 is represented by (x, y) colour values at a pixel, the Hue distance maps can be rewritten as follows.

$$h(c_{x,y}) = \exp\left[-\frac{1}{2}(c_{x,y} - \mu_{x,y})^T \sum^{-1} (c_{x,y} - \mu_{x.y})\right] \tag{2.5.3}$$

where $c_{x,y} = (x, y)$ and $\mu_{x,y} = (\mu_x, \mu_y)$. Figure 2.12 shows a road image with green sign board and its corresponding green conspicuity map. In this map, pixels with green colour have larger hue responses than any other chromatic pixels do. This conspicuity map shifts attentive focus on green traffic signs after it is weighted and linearly integrated into a final saliency map according to the modified FIT framework in Figure 2.9.



(a) Sample Image **(b)** Green Conspicuity Map

Figure 2.12: Sample Feature Map

Experiments & Discussion

The modified FIT framework for traffic-sign detection is tested on the UNMC Autovision database [104] with 119 images of driving scenes. This experiment aims to com-

pare performances between default FIT framework and modified FIT framework with top-down bias for traffic-sign detection. Visual results of the experiment are shown for five samples in Figures 2.13-2.17. Traffic signs cannot be detected in Figures 2.13a-2.17a if only three default channels of Itti toolbox: intensity, orientation, and (general) colour are utilised. In other words, general FIT saliency framework is not equipped for traffic sign detection. Figure 2.13b-2.17b show detection results if only six traffic sign colour channels, in the top-down bias part of the modified framework shown in the Figure 2.9, are used for computing saliency maps. If all nine channels are utilised simultaneously, visual results of traffic sign detections in five samples are shown in Figures 2.13c-2.17c.

Table 2.3: Simulation Results (N_s : $detection - rate$)

Method	**Option 1**	**Option 2**	**Option 3**
N_s	23/119	100/119	101/119

As Figures 2.13-2.17 show performance of the modified FIT framework with only default channels (Option 1), only traffic-sign colour channels (Option 2), or their combination (Option 3), Table 2.3 shows quantitative results according to these three options. In this table, N_s is a detection rate which is the ration between number of images where traffic-signs are successfully detected and the total number of images. According to the Table 2.3, option 1 gives the worst detection rate (23/110) since default channels in Option 1 is often distracted by color, intensity, orientation of other objects than traffic signs in natural road scenes. The other Options 2 & 3 (only traffic sign colour channels & all channels) give much better results, (100/119) and (101/119). In most of the testing images, traffic sign images are detected; moreover, even small traffic-signs can be found, as shown in the Figures 2.13-2.17. As Option 3 needs more computation power and processing time than Option 2 without significant improvement in detection rate, the modified FIT framework with traffic-sign colour channels alone proves to be sufficient for detection purpose. In brief, both visual and quantitative results show an advantage of top-down bias features, six traffic-sign colours, in the modified FIT model for detection. It proves that FIT saliency framework with suitable modification, a top-down bias, is useful for detection of specific objects rather than only general interesting points.

2.5.2 Frequency-based Saliency Methods for Lane-mark Detection

In the previous Sub-Section, a modified FIT sailency method with colour-based top-down bias help detect traffic-signs on road effectively. This Sub-Section shows that even without modification saliency methods can be useful for a certain computer vision application as well. Two frequency-based saliency methods, Phase Fourier Transform (PFT) and Quadrature Phase Fourier Transform (QPFT), are utilised for demonstrating



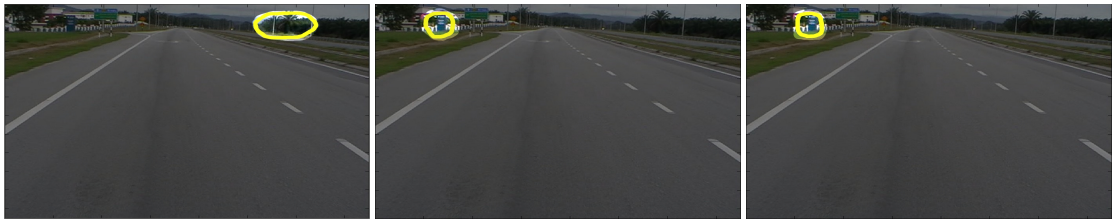
(a) Default Features **(b)** Traffic Sign Features **(c)** Both Features

Figure 2.13: Sample Image 1



(a) Default Features **(b)** Traffic Sign Features **(c)** Both Features

Figure 2.14: Sample Image 2



(a) Default Features **(b)** Traffic Sign Features **(c)** Both Features

Figure 2.15: Sample Image 3



(a) Default Features **(b)** Traffic Sign Features **(c)** Both Features

Figure 2.16: Sample Image 4



(a) Default Features **(b)** Traffic Sign Features **(c)** Both Features

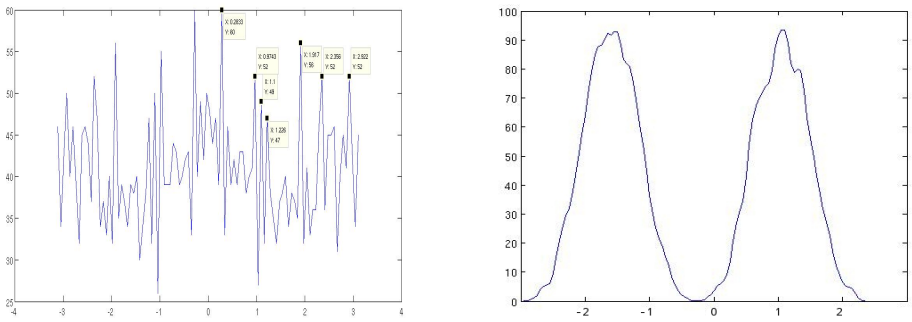
Figure 2.17: Sample Image 5

their usefulness on lane-mark detection. This proposal bases on following observations of lane-marks on roads. Big white or yellow lane marks obviously pop out as the most attentive items on grey roads. Usage of frequency-based saliency technique can quickly provide locations of the lane-marks as images are converted and processed in frequency domain.

Frequency-based Saliency (PFT or PQFT) Saliency can be seen as sudden disruption of pixel intensity spatially and temporally. In signal processing, these disruption can be represented under by large variance of amplitudes or phase. Guo *et al.*[2] shows that variances of phase is much more significant, meaningful and robust than those of amplitude. Then, their work has inspired the early idea of saliency frameworks in Fourier domain, Phase Fourier Transform (PFT) and Phase Quaternion Fourier Transform (PQFT), which are briefly summarized in the Sub-Section 2.4.2. While PQFT and PFT methods are designed for general object detection [2], they are used for lane-marks detection purpose in this Sub-Section.

Lane-marks as Saliency Object Lane marks are purposefully designed to attract drivers' attention in any possible weather conditions. Therefore, it can be assumed that lane-marks stand out of scenes in front of drivers most of the time. In the day time, lane-marks are painted in white or orange, attentive colours comparing to its surrounding context, grey roads. In the night time, lane-marks strongly react to other light sources; therefore, it definitely stands out from dark roads as well. There are usually changes from low-intensity regions, grey road regions, to high intensity regions, lane-mark regions. It causes disruptive changes in phase of Fourier spectrum, as shown in Figures 2.18a and 2.18b. It is noted that phase spectrums of PQFT, as shown in Figure 2.18a , is generally smoother than PQFT 2.18b since PQFT utilises all three colour as well as difference of intensity between consecutive video frames while PFT only utilises intensity of a single frame. Regardless of PFT or PQFT, these disruptive phase changes point out locations of salient objects [105]; in other words, identifying these sudden changes in phases means finding positions of interested lane-marks on road scenes. That is the reason why PFT and PQFT are useful in lane-mark extraction. The lane-mark extraction by PFT and PQFT are tested on several images and videos from UNMC Autovision Database, and results of PFT and PQFT lane-mark extraction are shown in Figures 2.19, 2.20 respectively. PQFT performs better than PFT for detection of general objects [106]. Visual results of six test samples among 96 scenes in UNMC Auto database [107] in Figures 2.19 and 2.20 show that PQFT performs better than PFT in lane-mark extractions as well. Comparison between Figures 2.19a-2.19f and 2.20a-2.20b shows that PQFT gen-

erates better boundaries for lane-marks than PFT does. Therefore, PQFT gives more accurate lane-mark extraction than PFT does.



(a) PFT **(b)** PQFT

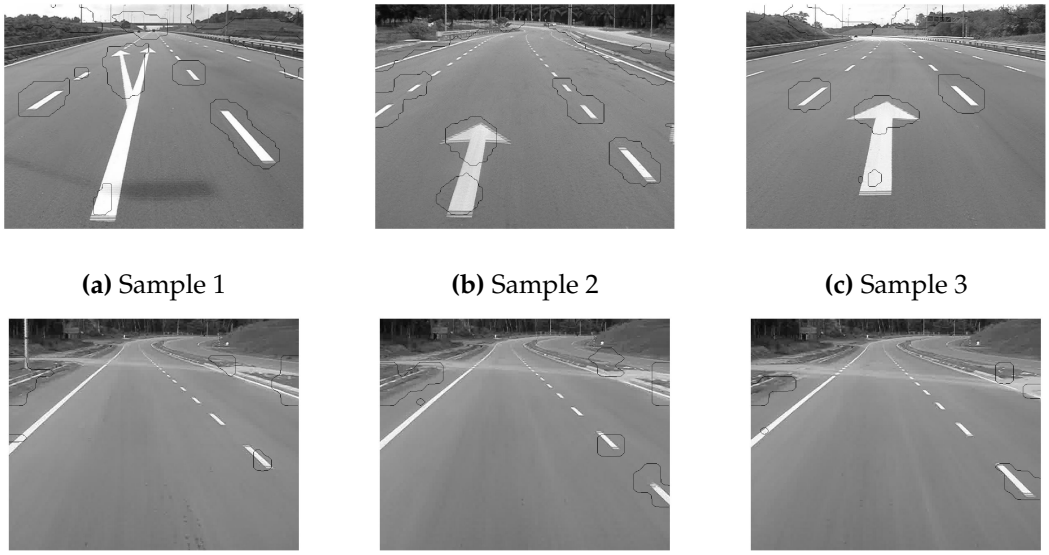
Figure 2.18: Phase Spectrum of Frequency-based Saliency Methods

Mechanisms behind PFT and PQFT are simple; therefore, they can be straightforwardly implemented by MATLAB Computer Vision System and DSP System Toolbox for real-time application on Texas Instrument DSP DM642. Graphical design of PFT and PQFT models are shown in Figures 2.21 and 2.22.

In brief, frequency-based saliency algorithms PFT and PQFT are useful for lane-marks detection. These saliency algorithms detect sudden disruption of phase values when images are converted to spectral domain from spatial one by Fourier Transform. These changes in phase values correspond to locations of interested subjects on scenes; in other words, lane-marks on road scenes. If saliency methods, PFT and PQFT saliency, are implemented properly in DSP platform, they can solve a specific real-time computer vision problem, lane-mark detection, even without any modification or top-down bias.

2.6 SUMMARY

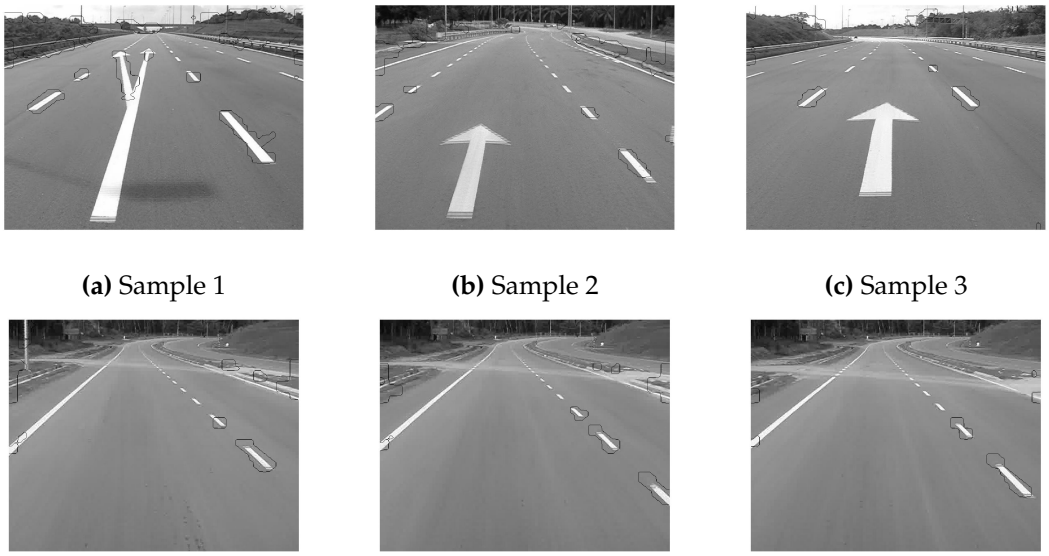
In a nutshell, the chapter gives an introduction to what saliency is about, how it should be modelled, which aspects need considering in modelling saliency, what approaches and theories are available. As visual saliency is proposed for numerically measuring how significant a feature pops out from a scene visually, its modelling involves many visual input features. The mentioned features are purposely designed to simulate centre-surround phenomenon, a contrast / difference in values between central and surrounding regions. Then experimental constrains, data evaluation processes for experiments with eye-fixation are described before signifiant computational saliency models in literature are discussed in details. The available theories and computational



(a) Sample 1 **(b)** Sample 2 **(c)** Sample 3

(d) Sample 4 **(e)** Sample 5 **(f)** Sample 6

Figure 2.19: PFT Lane-mark Extraction



(a) Sample 1 **(b)** Sample 2 **(c)** Sample 3

(d) Sample 4 **(e)** Sample 5 **(f)** Sample 6

Figure 2.20: Result 4

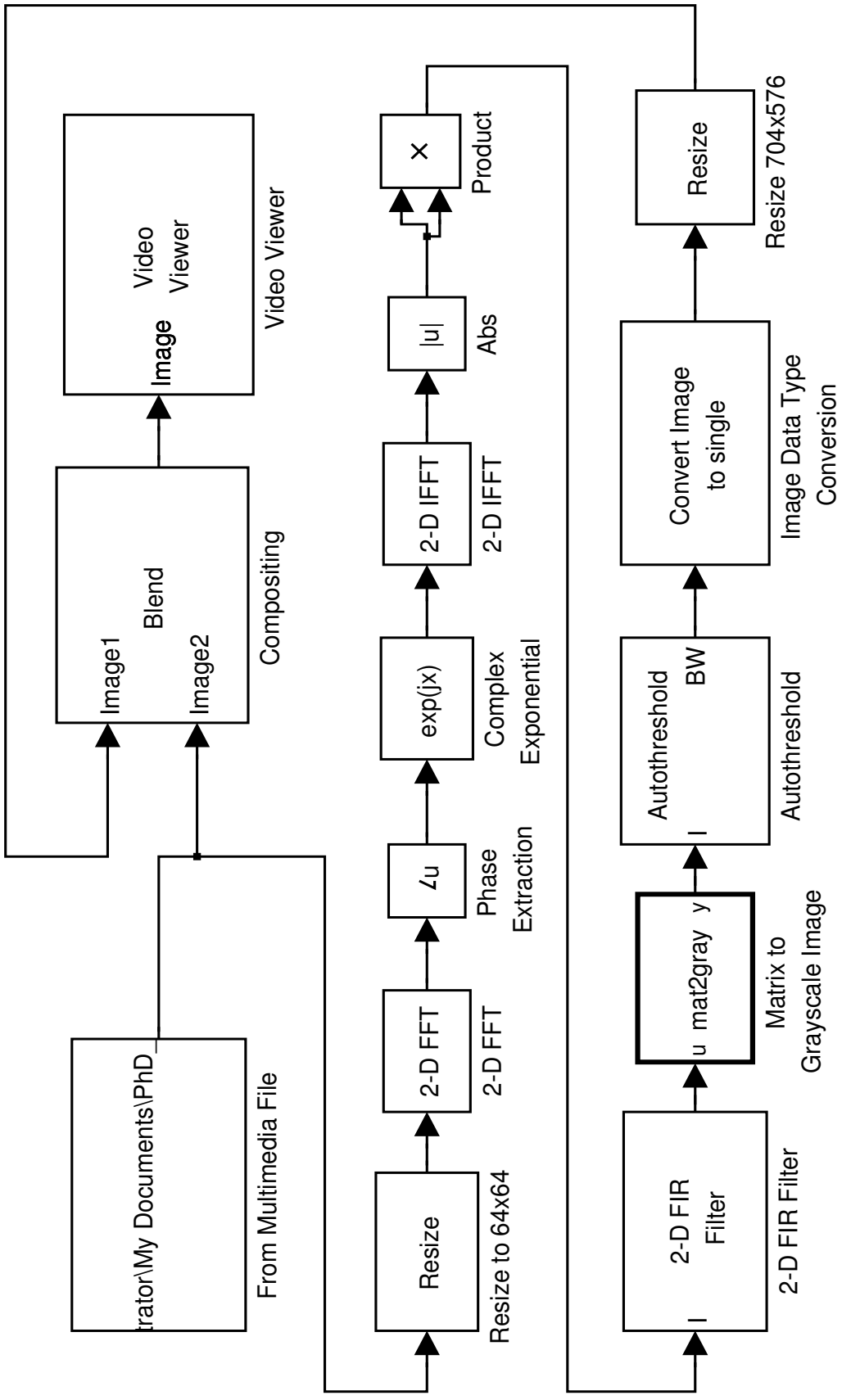


Figure 2.21: PFT Visual Saliency Method

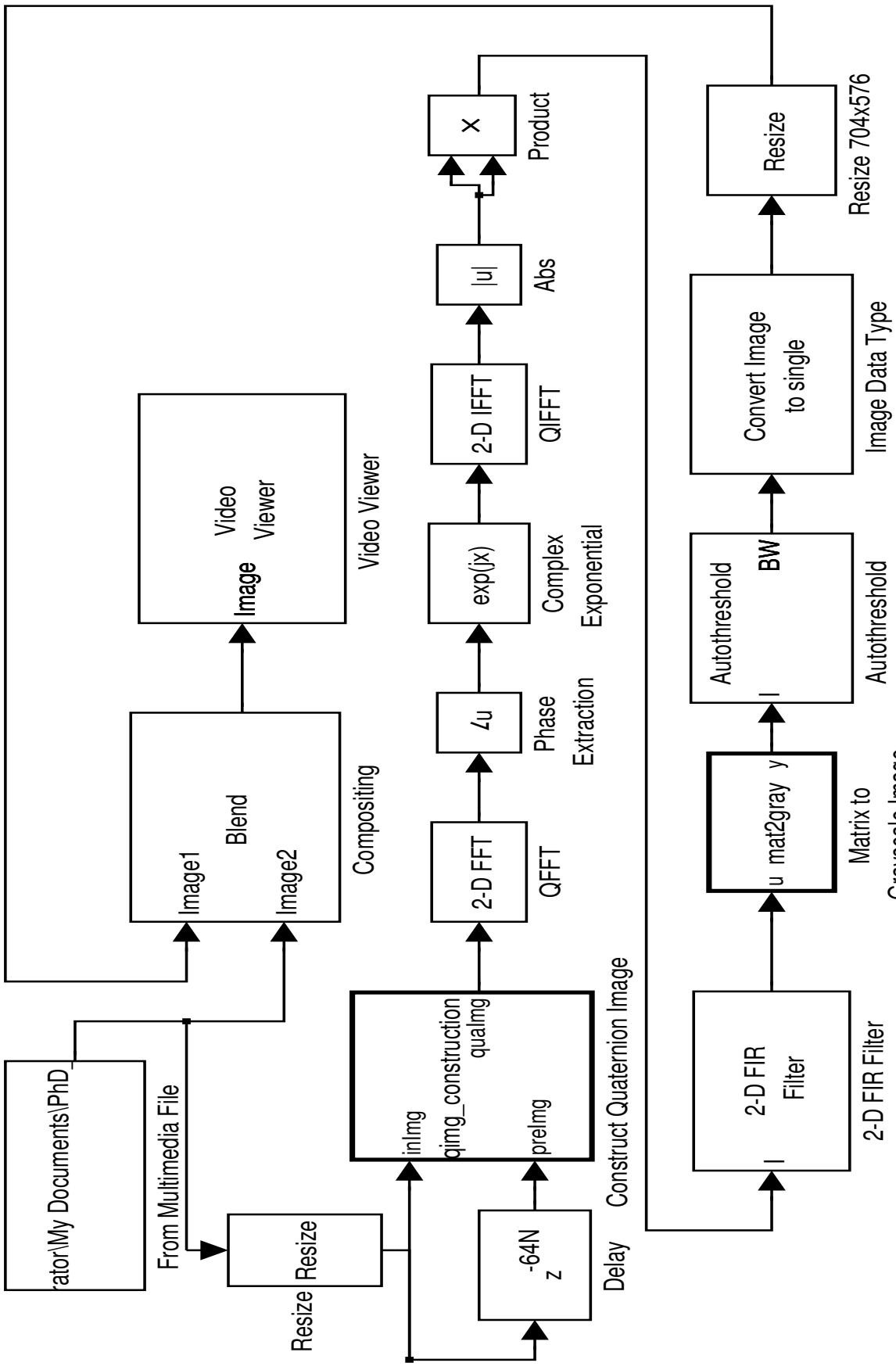


Figure 2.22: PQFT Visual Saliency Method

approaches are summarized into three main classes of saliency methods: feature-based saliency, phase-based saliency and information-based saliency. In each class, a few particular models are chosen to be described in details to explain core ideas of each saliency approach. Among three classes, information-based saliency models are especially focused since it directly motivates research topics of this thesis. Inspired by information theory, some saliency methods directly use information measurement concepts like local entropy, self-information in computation of saliency values, or indirectly a saliency method can utilise discriminant power for modelling visual attention and search with advantages of coping with wide range of attention related phenomena. Thoroughly, the free energy provides a comprehensive concept, covering all previous methods of information measurement and neurological structures of visual cells. Besides discussion of saliency concepts and computation, direct applications of saliency algorithms as solutions for practical computer vision problems are described as well. In these applications, saliency proves to solve practical problems in main processes rather than just extract interest points or regions in pre-processes.

CHAPTER 3

Fast non-parametric entropy-based saliency (ENT)

This chapter presents a formulation for bottom-up visual saliency as centre surround conditional entropy can be formulated with a fast and efficient technique for the saliency computation. It is shown that the new saliency formulation is consistent with self-information based saliency, decision-theoretic saliency and Bayesian definition of surprises but also faces the same significant computational challenge of estimating probability density in very high dimensional spaces with limited samples. The chapter proposes a fast and efficient non-parametric method to make the practical implementation of these types of saliency maps possible. After aligning pixels from the centre and surround regions and treating their location coordinates as random variables, a k-d partitioning method is employed to efficiently estimate the centre surround conditional entropy. Experimental results on two publicly available eye tracking still image databases show that the new technique is competitive with state of the art bottom-up saliency computational methods. The methods are extended for computing spatio-temporal visual saliency of video and evaluate the bottom-up saliency against eye tracking data on a video taken onboard a moving vehicle with the driver's eye being tracked by a head mounted eye-tracker.

3.1 INTRODUCTION

In this chapter , the centre-surround principle of bottom-up visual saliency is directly formulated as conditional entropy and Kullback-Leibler divergence. These quantities measure novel information (or uncertainty, informativeness, or surprise) of the centre given surrounds are known in advance. Though this formulation appears in previous

saliency models, e.g. self-information [4], mutual information [108] and Bayesian surprise [109], the success of these models is limited due to their complexity and heavy computational requirement due to the estimation of probability density functions in high dimensional spaces with limited samples. Meanwhile, a contribution of this chapter is formulating saliency as conditional entropy and Kullback-Leibler divergence in a direct and intuitive interpretation of centre-surround mechanism as well as facilitating an easy extension to spatio-temporal saliency. Besides intuitive interpretation of saliency, another contribution is adapting a fast and practical solution for estimating conditional entropy and Kullback-Leibler divergence. The fast algorithm of entropy estimation is inspired by non-parametric multidimensional k-d tree entropy estimation [21]. This implementation provides computationally tractable and feasible solutions for real-time applications such as saliency detection in videos. Experimental results are carried out on several publicly available eye-tracking image databases [4, 5] and a video taken on-board a car with driver's eye movements being tracked with a head-mounted eye tracker. They demonstrate effectiveness of proposed methods as well as compare the proposals with other state-of-the-art techniques.

3.2 FAST ENTROPY-BASED SALIENCY CONCEPT

Let $I_c(x,y)$ be an image patch at location (x,y) and $I_{sr}(x,y)$ its surrounding regions. The conditional entropy of the center given its surround can be defined as

$$\boldsymbol{H}(I_c(x,y)|I_{sr}(x,y)) = \boldsymbol{H}(I_c(x,y), I_{sr}(x,y)) - \boldsymbol{H}(I_{sr}(x,y)) \tag{3.2.1}$$

and can be further defined in terms of joint and marginal probabilities

$$\boldsymbol{H}(I_c(x,y)|I_{sr}(x,y)) = \sum_{\substack{I_c(x,y)\in\boldsymbol{I} \\ I_{sr}(x,y)\in\boldsymbol{I}}} p(I_c(x,y), I_{sr}(x,y)) \log \frac{p(I_{sr}(x,y))}{p(I_c(x,y), I_{sr}(x,y))} \tag{3.2.2}$$

, where $\boldsymbol{H}$ represents entropy and $p()$ is a probability distribution function (PDF). The conditional entropy $\boldsymbol{H}(I_c(x,y)|I_{sr}(x,y))$ can be understood in a number of ways. From a coding or information theory's perspective, it will take $\boldsymbol{H}(I_c(x,y), I_{sr}(x,y))$ bits to code the centre and its surrounds together. However, if the surround $I_{sr}(x,y)$ is known in advance, $\boldsymbol{H}(I_{sr}(x,y))$ bits of information is determined. Then, the conditional entropy measures the remaining bits needed to code the centre. From an uncertainty or informativeness point of view, the conditional entropy measures the remaining uncertainty of the centre once its surrounds are known, or the amount of information of the

centre given the knowledge of its surrounds. This information can be formulated as conditional entropy (CON) , Equation 3.2.3, or Kullback-Leibler divergence (KLD) as a measure of saliency, Equation 3.2.4, as follows .

$$\boldsymbol{S}(x,y) = \boldsymbol{H}(I_c(x,y)|I_{sr}(x,y)) \tag{3.2.3}$$

$$\boldsymbol{S}(x,y) = \boldsymbol{D}(I_c(x,y)|I_{sr}(x,y)) \tag{3.2.4}$$

, where $\boldsymbol{S}$ is a saliency value at a location (x,y) , $\boldsymbol{H}(I_c(x,y)|I_{sr}(x,y))$ is the conditional entropy and $\boldsymbol{D}(I_c(x,y)|I_{sr}(x,y))$ is Kullback-Leibler Divergence. The definition of saliency in Equations 3.2.3, 3.2.4 and 3.2.2 is consistent with a number of definitions in the literature including self-information [4], surprise [109] and decision theoretic saliency [108]. The self-information saliency of [4] measures the self-information of $I_c(x,y)$ in the context of its surrounds, $-\log(p(I_c(x,y)))$. If $I_c(x,y)$ is a common patch within the image, then $p(I_c(x,y))$ is large, $-\log(p(I_c(x,y)))$ will be small, hence the saliency will be small. $\boldsymbol{S}(x,y)$ in the Equations 3.2.3 and 3.2.4 has the same property, that is, if the center and its surrounds are very similar, then $\boldsymbol{S}(x,y)$ will be small and conversely, if they are very different, it will be large. The surprise measure of [109] can be re-written as

$$\boldsymbol{S} = \sum_{\substack{I_c(x,y)\in\boldsymbol{I} \\ I_{sr}(x,y)\in\boldsymbol{I}}} p(I_{sr}(x,y))log\frac{p(I_{sr}(x,y))}{p(I_{sr}(x,y)|I_c(x,y))} \tag{3.2.5}$$

Here, the surrounds $I_{sr}(x,y))$ can be interpreted as the model or background information and the center $I_c(x,y)$ as the new observation data. Comparing to our formulated definition of center-surround entropy in terms of Kullback-Leibler divergence,

$$\boldsymbol{D}(I_c(x,y)|I_{sr}(x,y)) = \sum_{\substack{I_c(x,y)\in\boldsymbol{I} \\ I_{sr}(x,y)\in\boldsymbol{I}}} p(I_c(x,y))log\frac{p(I_{sr}(x,y))}{p(I_c(x,y))} \tag{3.2.6}$$

There is a similarity between two types of saliency measurements defined by equations 3.2.6 and 3.2.5, so are their responses to input data. Again, the surprise measure will be small when the center and surround are similar and large when they are different. The decision theoretic discriminant saliency of [108] boils down to the computation of the mutual information between the center and its surround, and the mutual information and the conditional entropy has the following relation.

$$\boldsymbol{MI}(I_c(x,y), I_{sr}(x,y)) = \boldsymbol{H}(I_c(x,y)) - \boldsymbol{H}(I_c(x,y)|I_{sr}(x,y)) \tag{3.2.7}$$



Figure 3.1: Medium Band Filter Flow Chart

$\boldsymbol{MI}(I_c(x,y), I_{sr}(x,y))$ is the amount of uncertainty of the center $I_c(x,y)$ that is removed when its surrounds $I_{sr}(x,y)$ are known. One way to understand MI is that it measures how much the surround can tell about the center which is again consistent with the conditional entropy. A large mutual information means the surround can tell a lot about the center hence the saliency will be low, so will be the centre surround conditional entropy.

For all these definitions of saliency measure, there is a fundamental challenge - practical implementation. As all involve the estimation of probability density functions in very high dimensional spaces with limited data samples, various simplification processes have to be used, e.g., [4] employed independent component analysis and [108] assumed a parametric Generalized Gaussian Distribution (GGD) model. The next section introduces a fast non-parametric method.

3.3 FAST ENTROPY-BASED SALIENCY (ENT) MODEL

Visual data have excessive amount of information, but only some of them is useful for forming saliency maps. Itti *et al.*[1] used low-level features of intensity, colour and orientation to build several conspicuity maps which are combined in a linear fashion to generate a saliency map. In the discriminant saliency map approach, Gao *et al.*[108] used wavelet and Gabor filters to extract band-pass features and modeled these features using parametric GGD to estimate the mutual information between the centre and surround. In this chapter , features from medium frequencies are used since studies have shown that mid-band frequency data globally allow the best prediction of attention for many categories of scenes [20]. Figure 3.1 shows a step-by-step illustration of mid-band filtering used in this paper. In the first step, a 9/7 Cohen-Daubechies-Feauveau (CDF) wavelet[110] is used to analyse the image into three different frequency bands. After isolating different frequency bands of the visual data, the lowest and highest bandpass of the wavelet analysis are removed in step 2. The remaining components are converted back to the image domain by the inverse of the 9/7 CDF wavelet in step 3. The lowest sub-band is filtered out in order to remove global trend

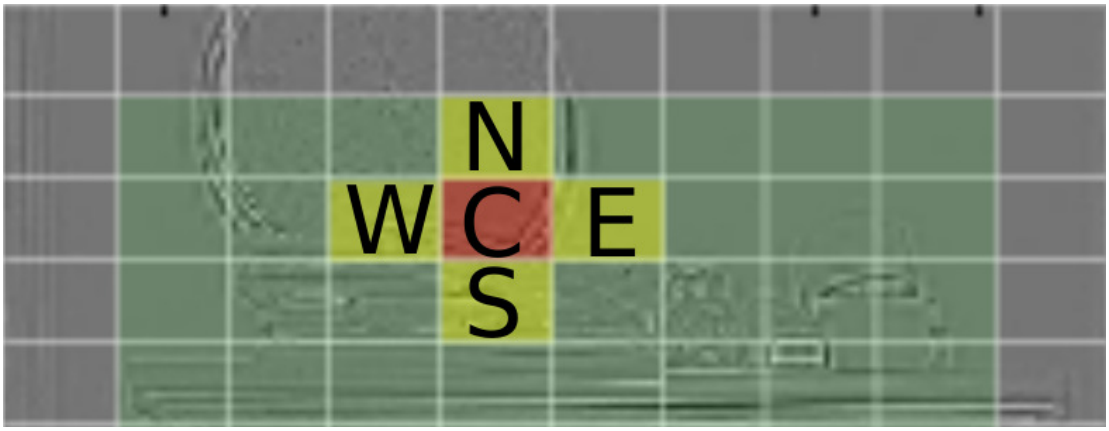


Figure 3.2: Effective Saliency Region

of signals or images while the proposed saliency method mainly depends on centre-surround mechanism or local difference between central and surround patches. Elimination of wavelet coefficients in the first level, highest spectrum of frequency, help to remove noise. These two steps named as Medium Subband Filter (MSF) is proved to be biologically plausible. Urban *et al.*[20] carry out several experiments on correlation between eye fixation data and image features of different frequency bands and concluded that human visual performance corresponds well with image features from wavelet frequency band L2 to L4 [20]. In addition to that good relativity between psychological data and MSF features, another advantage of MSF is cleaning data since it removes global trend and noise signals to which the proposed method of entropy estimation is very sensible. More details about the preprocessing steps are mentioned in Section 3.5. After medium frequency filtering, the outputs of the step 3 are divided in 8x8 patches as shown in Figure 3.2. Although other patch sizes are possible, 8x8 patches were used in this paper. From those filtered data in patches, saliency values can be estimated as conditional entropy (CON) of central data given surrounding data, Equation 3.3.1, or statistical Kullback-Leibler divergence (KLD) of data between centre and surround patch, Equation 3.3.4 .

First, the saliency as the conditional entropy of each centre patch (**C**) is considered given its four surrounding patches (**N**, **S**, **W**, and **E**) as shown in Figure 3.2. The conditional entropy based saliency is computed as

$$S(\mathbf{C}) = H(\mathbf{C}|(\mathbf{N},\mathbf{S},\mathbf{W},\mathbf{E})) = H(\mathbf{C},\mathbf{N},\mathbf{S},\mathbf{W},\mathbf{E}) - H(\mathbf{N},\mathbf{S},\mathbf{W},\mathbf{E}) \tag{3.3.1}$$

where S and H are a saliency value and entropy value. Estimating the two joint entropies on the right-hand side of the Equation 3.3.1 is challenging because of the high dimensionality of the data. To get round the problem, coordinates of pixels are treated

as the random variables [111] and the Equation 3.3.1 is approximated as.

$$\begin{aligned} S(\mathbf{C}) =& H(c(x,y), n(x,y), s(x,y), w(x,y), e(x,y)) \\ &- H(n(x,y), s(x,y), w(x,y), e(x,y)) \end{aligned} \tag{3.3.2}$$

where $c(x,y), n(x,y), s(x,y), w(x,y), e(x,y)$ are respective pixels from the $\mathbf{C}, \mathbf{N}, \mathbf{S}, \mathbf{W}, \mathbf{E}$ patches at the same reference location (*x*,*y*).

The problem is treated as drawing samples from (*x*,*y*) in order to approximate the conditional entropy. With the formulation of Equation 3.3.2, the problem can be simplified as estimating the entropies in the 4-D and 5-D spaces with a total of 8x8 = 64 samples. A similar technique is used by Stowell *et al.*[21] to achieve fast implementation of conditional entropy in the Equation 3.3.2. The technique is based on a k-d tree entropy estimation approach to partition the input data space $\Omega \in \Re^D$ into $A = \{A_j | j = 1, 2, \ldots, m\}$ with $A_i \bigcap A_j = \varnothing$ if $i \neq j$ and $\bigcup_j A_j = \Omega$. Let n_j be the number of samples in the cell A_j, $V(A_j)$ the volume of cell A_j, the total number of sample N, then the multidimensional joint entropy $\hat{H}$ can be estimated as

$$\hat{H} = \sum_{j=1}^{m} \frac{n_j}{N} log\left(\frac{N}{n_j} V(A_j)\right) \tag{3.3.3}$$

As well as conditional entropy, Kullback-Leibler divergence is another measure which suits into centre-surround difference computation. According to the equation 3.2.6, the saliency of central patch is its divergence in accordance with the surrounding reference. Then, the computational task is figuring out $p(I_{sr}(x,y))$ and $p(I_c(x,y))$, where $I_{sr}(x,y) = \{n(x,y), s(x,y), e(x,y), w(x,y)\}$ and $I_c(x,y) = \{c(x,y)\}$.

Once again, the technique proposed in [21] comes in handy. The k-d tree partition technique can be utilized to break data points into uniformly distributed cells where probability density function can be simply approximated with number of samples and its correspondent volumes. However, there is one distinguishing point between KDPEE usage in conditional entropy and estimation of KL divergence which is data ordering. In conditional entropy estimation $H(C|S)$, it does not matter that the joint entropy is computed as $H(C,S)$ or $H(S,C)$. However, that order has important meaning in KL divergence which is asymmetric information measurement $D(C||S) <> D(S||C)$. Hence, before performing k-d partition, data should be organized such that first four dimensions contains data of surrounding patches, and the fifth dimension contains data of central patch. That arrangement partitions surrounding data before any central data is done. In other words, surrounding information already gives reference to partitioning steps for central patches. The partition scheme ensures that number of samples

in central cells and surrounding cells are equal $n_j^{sr} = n_j^c$, so the ratio $\frac{p(I_{sr}(x,y))}{p(I_c(x,y))}$ is equal to $\frac{V(A_j^c)}{V(A_j^s)}$. Therefore, the multidimensional joint KL divergence $\hat{D}$, an approximation of $\boldsymbol{D}(I_c(x,y)|I_{sr}(x,y))$ in Equation 3.2.6, can be written as.

$$\hat{D} = \sum_{j=1}^{m} \frac{n_j}{N} log\left(\frac{V_c(A_j)}{V_{sr}(A_j)}\right) \tag{3.3.4}$$

The computational complexity of either conditional entropy or Kullback-Leibler approach is $\Theta\left(DNlogN\right)$ and the space complexity is $\Theta(DN)$ whereof $N = 64$ and $D = 5$ or $D = 4$ in these proposed methods. A lower limit of a suitable sample size is $N \geq 2^D$, $2^5 = 32$ and $2^4 = 16$; therefore, the number of considered pixels, $N = 64$, satisfies the lower-limit of samples size requirement.

Conditional entropy (CON) and KL diverge saliency (KLD) formulations are similar in principles, measurements of centre-surround information, and computational methods, usage of kd tree partitioning. However, KL divergence requires less computation and produces smoother visual saliency maps. Two KDPEE are required for conditional entropy measurement meanwhile KL divergence only needs one. Therefore, the later approach is essentially twice as fast as the former. In addition, KL divergence reduces the effect of bias in information estimation which heavily affects saliency maps generated by conditional entropy, differences between two entropy estimation. Therefore, it produces much smoother or blurred saliency maps which more likely match distributions of human fixations [112] .

Besides better performance, KL divergence possesses interesting theoretical points which help to measure bias in information estimation while conditional entropy does not clearly account for it. According to Gibbs' inequality, Kullback-Leibler divergence must be positive; therefore, any negative KL divergence is a result of estimation bias. The bias is mainly due to samples shortage in high-dimensional data estimation; in other words, the bias ratio, number of biased estimation over the total number of estimation, depends on the size of patches, amount of available data for estimation. It paves a way for identifying data size which minimizes possible bias. For example, in 544x720 images, if the odd patch size, n, is varied between 7 and 21, the correspondent bias ratio are as follows.

Table 3.1: Bias ratio vs. patch size

Patch size	*7x7*	*9x9*	*11x11*	*13x13*	*15x15*	*17x17*	*19x19*	*21x21*
Error ratio	0.0352	0.0250	0.0116	0.0046	0.0014	0.0002	0.0000	0.0000

Results in Table 3.1 points out that a size of patches may need to reach 19 or 21 to totally eliminate bias in KL divergence estimation. Though 7x7 patch size is responsible

for 3.52% more bias estimation points than 21x21 patch size, it does not affect relative saliency value comparison between two points as long as they are estimated on the same number of data. Therefore, it does not cause any significant differences in normalized saliency maps. Choices of size patches do not cause much difference visually and numerically in normalized saliency maps, but it does affect performance in terms of speed. Obviously the more data are input, the more processing needs to be done. In later experiments, 7x7 patches are mainly utilized, since it provides both reasonable saliency maps and computational efficiency.

3.4 ENTROPY-BASED SALIENCY FOR VIDEO

The techniques described in 3.2 and 3.3 can be easily extended to 3D and to compute spatio-temporal saliency maps (STSM) for videos. Let $I_c(x,y,t)$ be the central patch at frame t, and $I_{ssr}(x,y,t)$ its spatial surround defined similarly as before, 4-neighbouring North, South, East and West patches. While $I_{tsr}(x,y,t)$, the temporal surroundings are central patches of past consecutive temporal frames, $\{I_c(x,y,t-1), I_c(x,y,t-2), ..., I_c(x,y,t-n)\}$. To simplify notation, we will drop the spatial coordinates without causing confusion. The choice of surrounding data leads to reasonable assumption that there is independence between spatial and temporal contexts [113], then we can define spatio-temporal saliency in term of conditional entropy and Kullback-Leibler divergence as

$$S_{st}(t) = \boldsymbol{H}(I_c(t) \mid I_{ssr}(t)) + \boldsymbol{H}(I_c(t) \mid I_{tsr}(t)) \tag{3.4.1}$$

$$S_{st}(t) = \boldsymbol{D}(I_c(t) \mid I_{ssr}(t)) + \boldsymbol{D}(I_c(t) \mid I_{tsr}(t)) \tag{3.4.2}$$

The two conditional entropy or KL divergence can be estimated similarly using the technique mentioned in the section 3.3. Moreover, the spatio-temporal saliency method, the equation 3.4.1, is utilised to process videos taken in the perspective of drivers' eyes with locations being recorded with a head-mounted eye tracker. The video and the eye tracker are synchronized so the driver's fixation points can be simultaneously shown in the video. Then, a saliency map is generated for each video frame accordingly, and its correlation with recorded fixation points provides a clue in how well saliency maps would predict driver's attention in real life. A sample scene and correspondent saliency map corresponding to different saliency methods are shown in Figure 3.3. Though results of spatial-temporal saliency seem to fit into context of images, the method has not been well optimized and analysed yet. Figure 3.4, from left to right, shows original



Figure 3.3: Samples and Saliency Maps of ITT, GBVS, PFT, and PQFT, and ENT methods (left-to-right top-to-bottom order). Please see the supplementary materials for sample sections of the video containing the original scene with the eye fixation marker overlayed on it and saliency maps of various methods.

image, spatial and temporal saliency maps according to estimation of conditional entropy or KL divergence values. In this figure, MSF is abbreviation of utilised Medium Sub-band Feature; Spatial Saliency Map (SSM) and Temporal Saliency Map (TSM) represent spatial and temporal saliency maps respectively; meanwhile, CON and KLD notify Conditional Entropy, the Equation 3.3.2 and Kullback-Leibler divergence, the Equation 3.3.4, as a strategy of information estimation. First, all saliency maps regardless of approaches (SSM / TSM or CON / KLD) identify edges as high saliency regions. Intuitively, edge regions of images indicates sharp change in intensity values between pixels as well as large statistical difference between neighbouring blocks in both spatial and temporal domain. Therefore, surrounding information of edge blocks does not provide much information for identification of central block statistical properties; then, large entropy difference of entropy occurs at these edge blocks. Secondly, visual results differentiate performance of KLD and CON approaches; KLD method generates smoother and blurred saliency map with less contrasting range; whereas, CON saliency map is much more sharp with larger range. However, both of them are sensitive to noise, or small change in high intensity area for example sky regions. Besides noise levels in KLD and CON maps, another observable point is high similarity between spatial and temporal saliency maps. The high correlation of pixels between successive frames results in a TSM with much similarity to a spatial saliency map (SSM). Current drawbacks of the proposed method are solely due to its sensitivity to noise and correlation between neighbouring data, and these problems can be solved by de-noising and decorrelation technique, introduced in the next section.

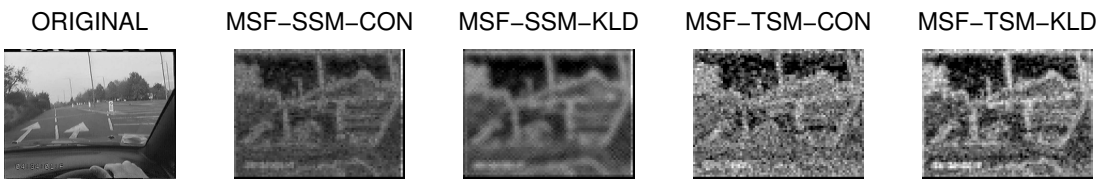


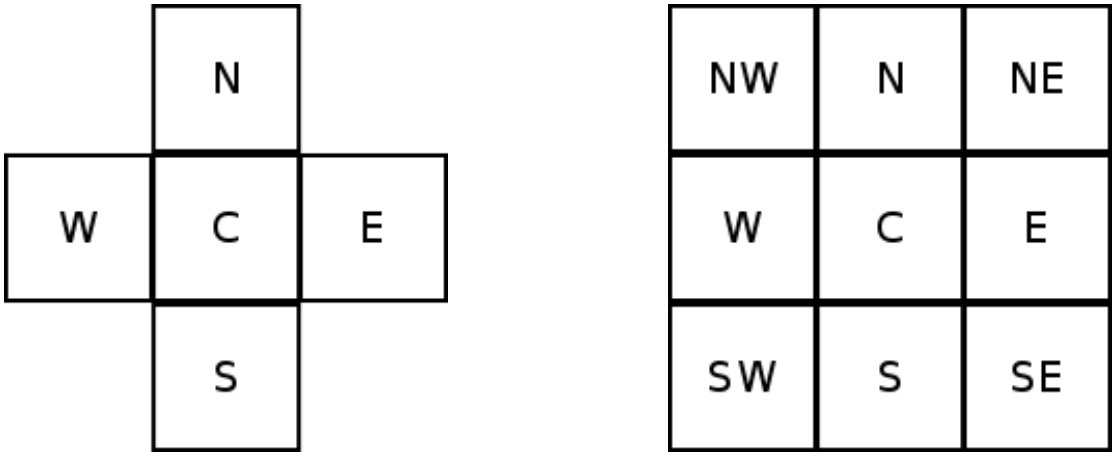
Figure 3.4: Standard Deviation of Each Sub-band in Camera-man Image

3.5 ENTROPY-BASED SALIENCY & DATA PREPROCESSING

As mentioned in the previous section, the method of fast entropy estimation empirically partition data into normal distributed groups by kd-tree methods and then computes entropy using a simplified entropy formula given the normality of data distribution. Since multi-dimensional data need to be dealt with, adaptive kd-tree is utilized instead of traditional kd-tree partitioning processes. When input data are multidimensional, the advantage of this adaptive process is allowing data to be split along different dimensionality at various levels. Then, instead of partitioning n-dimensional data all together, data can be split sequentially one dimension after another. In order to ensure close approximation of adaptive strategy to normal partition, both correlation of data and number of dimensions need to be as small as possible. Therefore in data preprocessing stage, a decorrelation process is crucial for accuracy of entropy estimation as well as the whole entropy-based saliency approach. Beside correlation between multidimensional data, its noise is another factor which sometimes severely affects the final saliency value, so noise removal techniques should be considered before any estimation is taken place.

3.5.1 Data Preprocessing - Dimension Reduction & Decorrelation

As mentioned in computing approach of spatial saliency map, information needs to be estimated from high dimensional dataset. In details, it is the amount of additional information for encoding a central patch data given 4-connected surrounding patches, as shown in Figure 3.2. In 4-connectivity case Figure 3.5a, CON first carries out estimation of 5-D dataset, (**C,N,S,W,E**), and 4-D dataset, (**N,S,W,E**). Then their differences are used as saliency values of CON saliency maps. KDPEE needs to estimate from a 5-D data; therefore, the number of pixels per a local patch needs to be at least 32 $(number\_of\_samples) > 2^5$. If a patch size is chosen as either 8x8 or 9x9, 64 and 81 pixels are more than sufficient for 4-connectivity neighbourhood. However, if 8-connected neighbouring patches, as shown in Figure 3.5b, are considered. KDPEE needs to estimates from 9-D dataset of (**C,N,S,W,E,NS,NW,SE,SW**) instead of 5-D dataset and 8-



(a) 4-connectivity **(b)** 8-connectivity

Figure 3.5: Surrounding neighbourhood connectivity

D dataset of (**N,S,W,E,NS,NW,SE,SW**) instead of 4-D dataset. Theoretically, at least $2^9 = 512$ pixels/patch would be needed; however such huge local patches are not available in experiments of 512x512 or 1024x1024 images. Insufficient of data prevents usage of 8-connectivity in computation of CON and KLD saliency maps. This insufficiency of pixels can be avoided if dimensionality of data is reduced. For example, if 9-D and 8-D data can be preprocessed and decorrelated into 5-D and 4-D data, it brings the amount of pixels down from 512 to 32 pixels per patch. However, data decorrelation or dimensional reduction also alters distribution of data which affects KDPEE's performance. Moreover, it might cause divergence in information estimation and noise in final saliency maps. This sub-Section aims to partly clarify effectiveness in utilisation of 8-connectivity data and dimension reduction method.

In image processing and computer vision research field, Principal Component Analysis (PCA) is well established as de-facto decorrelation tools for data processing because of its reliable theory and good practical performance for short signals of low-dimensional data. Besides decorrelation, through the analysis, data dimensions can be reduced by eliminating projected low-energy PCA coefficients. It comes in handy in this work since an amount of samples desperately limits the number of data dimensions which can be reliably estimated. Therefore, PCA is utilised for decorrelating and reducing dimensionality of a local patch, Figure 3.6. This figure presents distribution of energy across infinite PCA basis. Though infinity of horizontal axis, the energy values at the sixth basis and later are insignificant. Most of energy and information of patches exist in first few bases, as shown in Figure 3.6. Therefore, only first few PCA features, 4 or 5 first basis, occupies more than 99.8% of the total energy, can safely describe the pre-processed input data.

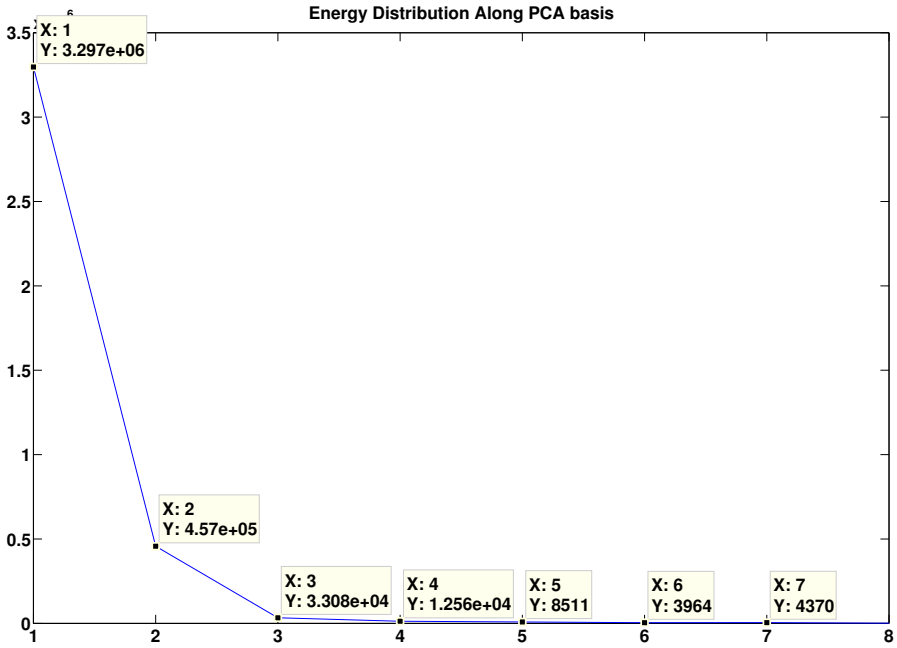


Figure 3.6: Energy Distribution according to PCA basis

An experiment is carried out to compare between spatial saliency maps generated by the proposed ENT-CON and ENT-KLD methods with/without decorrelation, Figure 3.7. On the first row from left to right, there are spatial saliency maps (SSM) generated by CON and KLD without PCA decorrelation and dimensional reduction. Meanwhile, the second row show saliency maps in that order but with decorrelation and dimensional reduction step (DEC) . The visual results in the Figure 3.7 show that decorrelation steps is not necessary for computation of spatial saliency method; moreover, it creates more noise in final saliency maps. Saliency values are collected at two points with following coordinates (328,311) (an edge point) and (467,57) (sky point). At edge points, saliency values are not much different between methods with or without decorrelation steps. However, the values are significant different on supposedly not salient sky point, and they tend to become bigger at those unimportant positions. Visually, decorrelated data generates more noise on the spatial saliency maps. It shows that 4-connecting neighbours without decorrelation provides better context for spatial saliency than 8-connecting neighbours with decorrelation. Therefore, decorrelation step is not necessary for spatial saliency map generation. Moreover, it takes quite a lot of computational effort because it takes averagely three times longer to process data with this decorrelation step. Noises of sasliency maps, generated with 8-connectivity and DEC steps, are due to changes of data distribution after DEC and sensitivity of KDPEE on such changes. DEC bases on PCA, which projects original data onto a basis to maximize an entropy of their distribution [103]. Visually, it spreads data out as much as possible. Though it benefits condensation of signal energy on first few bases, it makes entropy

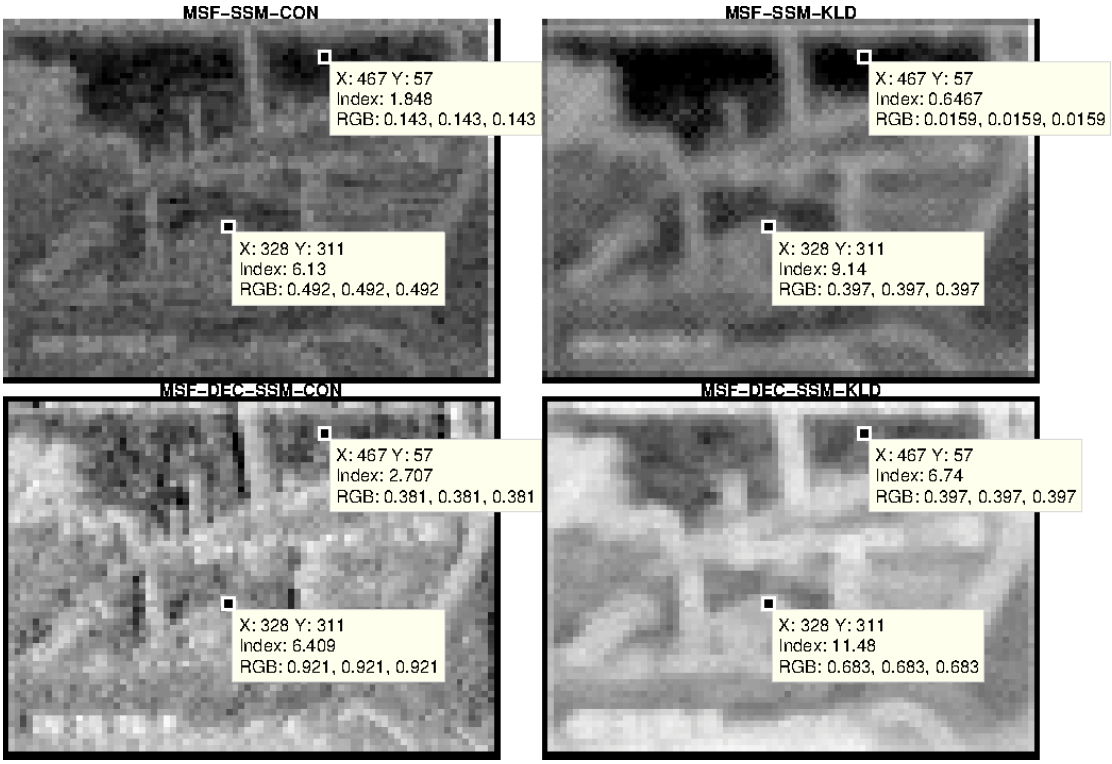


Figure 3.7: Spatial Saliency Maps by CON and KLD with (the bottom row) / without (the top row) DEC (decorrelation)

estimation of KDPEE difficult to converge. As a strategy of KDPEE information estimation is diving and conquering, it try to split data into homogeneous groups of pixels, whose values forms a uniform distribution. With overly expanded decorrelated data, KDPEE would probably struggle to form such homogeneous groups and converge in a well-approximated entropy value. This analysis shows that a specific decorrelation and dimensional reduction techniques needs to be developed to work with KDPEE. Otherwise, introduction of an arbitrary method, for instance DEC, for 8-connectivity introduces more noises into CON and KLD saliency maps.

3.5.2 Discrete Cosine Transform - Temporal Data Decorrelation

Though PCA is often used for decorrelating data and reduce data dimension, it is computationally expensive to carry out that analysis repeatedly. Moreover, temporal data need to be dealt with in this section, and Discrete Cosine Transform (DCT) has strong support as fast alternative for decorrelating temporal data instead of PCA in image and video compression. When there are several movements in the scene like two image sample images on relatively noisy background, first column in Figure 3.8. The temporal saliency maps generated by prior temporal approach - the middle column of Figure 3.8 - clearly show that current method can not cope with those noisy signals and

Algorithm 1 Medium Sub-band Frequency Filter - Temporal Data Decorrelation

Require: $input\_frames$
$msf3d\_filter\_decorrelation(input\_frames)$
$wavelet\_coefficients = waveletcdf97(input\_frames, number\_wavelet\_level)$;
$wavelet\_coefficients\{L1\} \leftarrow 0$; {Assign 0 to all L1 wavelet coefficients}
$wavelet\_coefficients\{DC\} \leftarrow 0$; {Assign 0 to all DC wavelet component}
$dct\_coefficients\_reduced = temporal\_decorrelation(wavelet\_coefficients)$;
$msf\_denoised\_feature =$
$inverse\_waveletcdf97(dct\_coefficients\_reduced, number\_wavelet\_level)$;
Ensure: $msf\_decorrelated\_feature$

their change over time. There is large temporal correlation between pixels at the same location of consecutive frames for both salient foreground features and insignificant background features. In addition, if a whole stack of consecutive patches of images without preprocessing is utilized for estimating temporal saliency, it partially covers spatial saliency as well since a whole spatial feature data are in use. Meanwhile, interest of temporal saliency method is emphasizing large movement or big difference in pixel values of video frames. Therefore, in order to decorrelate patch data as well as concentrate on motion features, DCT is employed. Besides its decorrelation mentioned before in video compression literature, its 1D computational processes emphasizes on difference of pixels at the same location but from different frames. Its coefficients eventually resembles changes of frames over time. After applying DCT analysis on MSF features of several frames, temporal saliency maps become much clearer and they actually highlight moving objects on a table tennis, right column of Figure 3.8. In addition, the largest coefficients of DCT correlation are removed from temporal saliency estimation in order to limit affects of spatial features on temporal saliency values and noise from background features. Beside removing of the largest component due to its spatial feature, a few smallest DCT coefficients are as well eliminated in order to reduce data dimensions and complexity while its general structure is still intact and meaningful. For example, 8-D data set, composed from 8 consecutive frames, might be simplified into 3-D by removing the first biggest and the last four smallest coefficients; then the dataset is much simplified and focused on frame differences. The whole temporal decorrelation process is summarized in pseudo-code Algorithms 1,2. Moreover, Figure 3.8 shows that KLD is better than CON when noise and complex texture are dominating background of sample images. However, generally temporal saliency maps are still plagued with noisy non-salient parts. Therefore, de-noising, another preprocessing data step, need integrated besides decorrelating process.

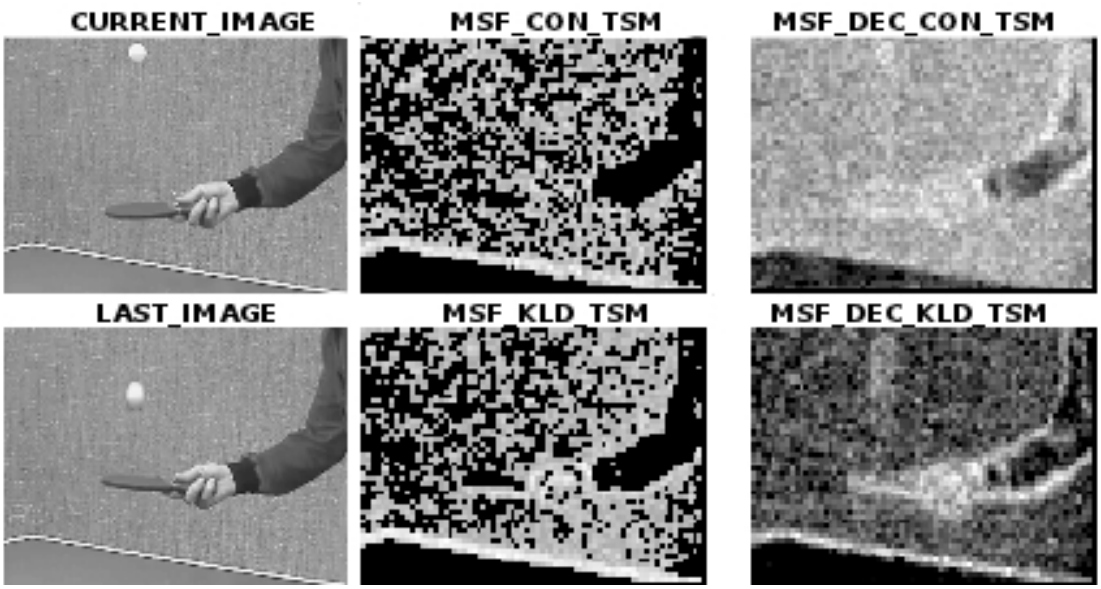


Figure 3.8: Temporal Saliency Map without / with correlation

3.5.3 Data Preprocessing - Image De-nosing

Spatial Saliency Map De-noising

Though in theory MSF features provides suitable data for KDPEE estimation due to its freedom from DC signals and level-1 wavelet features, noise does occur in medium-frequency subbands as well. Sometimes, they can cause significant drift in general performance of saliency methods or over/under-estimation of entropy-based salience values. Therefore, beside removal of the highest frequency components from wavelet analysis, further technique must be utilized to suppress unwanted background-noise in the rest of sub-band without eliminating entire frequency spectrum of images signals. Bivariate shrinkage functions Sendur *et al.*[114] are employed since its de-noising mechanism operates on wavelet domain, a compulsory step in the pre-processing stage. Briefly, its noise suppression technique based on correlation between parents and children wavelet coefficients which implies the wavelet coefficients have high possibility of belonging to salient edges feature if its parent coefficients are extracted from edge features and vice versa. In other words, bivariate shrinkage reserves coefficients which are salient across several scales or sub-band and eliminate noise features which are often only salient at a specific scale. Necessary steps for eliminating noise across scales in this proposal is detailed in pseudo-codes of Algorithm 3. De-noising effect on spatial saliency maps can be seen on spatial saliency maps of Figure 3.9. It shows saliency maps generated by CON (top row), KLD (bottom row) with de-noising affect (left column) and without denoising affect (right column). Its effects can be clearly observed on the bottom row, applying bivariate shrinkage before saliency value KLD estimation

Algorithm 2 Temporal Data Decorrelation & Dimension Reduction

Require: $wavelet\_coefficients$
{Temporal Data Decorrelation Step}
for $pixel\_counter = 1$ **to** $number\_of\_pixels$ **do**
  for $frame\_counter = 1$ **to** $number\_of\_frames$ **do**
    $pixel\_temporal\_data = input\_frames(pixel\_counter, frame\_counter)$;
  end for
  $dct\_coefficients(pixel\_counter, :) = dct(pixel\_temporal\_data, :)$;
end for
{Data Dimension Reduction Step}
for $pixel\_counter = 1$ **to** $number\_of\_pixels$ **do**
  for $dct\_basis\_counter = 1$ **to** $number\_of\_dct_basis$ **do**
    $basis\_energy(dct\_basis\_counter) +=$
    $\{dct\_coefficients(pixel\_counter, dct\_basis\_counter)\}^2$;
  end for
end for
{Sort Basis Energy in Descending Order}
$basis\_energy\_sorted\_index = sort(basis\_energy, ''descend'')$;
for $pixel\_counter = 1$ **to** $number\_of\_pixels$ **do**
  for $index\_counter = 2$ **to** $number\_dct\_basis/2$ **do**
    $dct\_coeffcients\_reduced(pixel\_counter, basis\_energy\_sorted\_index(index\_counter)) \leftarrow$
    $dct\_coefficients(pixel\_counter, basis\_energy\_sorted\_index(index\_counter))$;
  end for
end for
Ensure: $dct\_coefficients\_reduced$

technique generates smooth and nearly free noise spatial saliency maps. Especially, this de-noising technique incorporates well into the previous MSF feature extraction because it reuses wavelet CDF 9/7 analysed coefficients. Therefore, it does not require much additional computational effort. Due to its good performance and, that bivariate shrinkage might be extended to suppress noisy temporal saliency maps, Figure 3.8

Algorithm 3 Medium Subband Frequency Filter - Bivariate Shrinkage De-noising

Require: $input\_image, number\_wavelet\_level$
$wavelet\_coefficients = waveletcdf97(input\_image, number\_wavelet\_level)$;
$wavelet\_coefficients\{L1\} \leftarrow 0$;
$wavelet\_coefficients\{DC\} \leftarrow 0$;
$denoised\_wavelet\_coefficients =$
$bivariate\_shrinkage(wavelet\_coefficients, number\_wavelet\_level - 1)$;
$msf\_denoised\_feature = inverse\_waveletcdf97(input\_image, number\_wavelet\_level)$;
Ensure: $msf\_denoised\_feature$

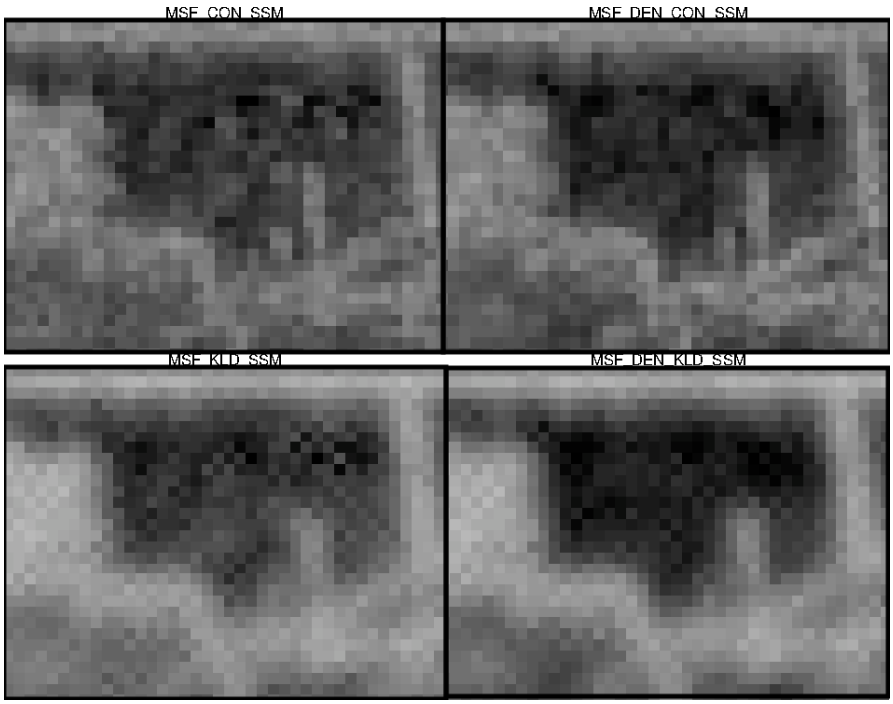


Figure 3.9: Spatial Saliency Map: (left) without denoising, (right) with denoising

Temporal Saliency Map De-noising

Previous sections have already elaborated about how bivariate shrinkage technique removes noise in spatial saliency map. As fore-mentioned, the first largest DCT basis and a few smallest basis in term of energy are eliminated to avoid spatial features and reduce data complexity. However, their saliency maps are still affected by noise which exist in the rest of DCT basis. Bivariate shrinkage is once again deployed to suppress noise on each DCT basis by using correlation between parent and children wavelet coefficients, pseudo-codes of Algorithm 4. Table tennis samples are again used to demonstrate how much noise can be removed from temporal saliency maps after bivariate shrinkage approach. Almost all background noise have been removed from the saliency maps as illustrated in the left-most column of Figure 3.10. Both CON and KLD successfully highlights supposedly saliency objects on the scene which are moving table tennis balls, and arms of players.

3.6 DISCUSSION & RESULTS

The previous sections have discussed rationales of entropy-based saliency method (ENT) with two information estimation strategies CON and KLD as well as necessary preprocessing steps. This section shows evaluation these proposed methods on publicly available eye tracking databases of Bruce *et al.*[4], Judd *et al.*[5], AUTOM, MSDR databases

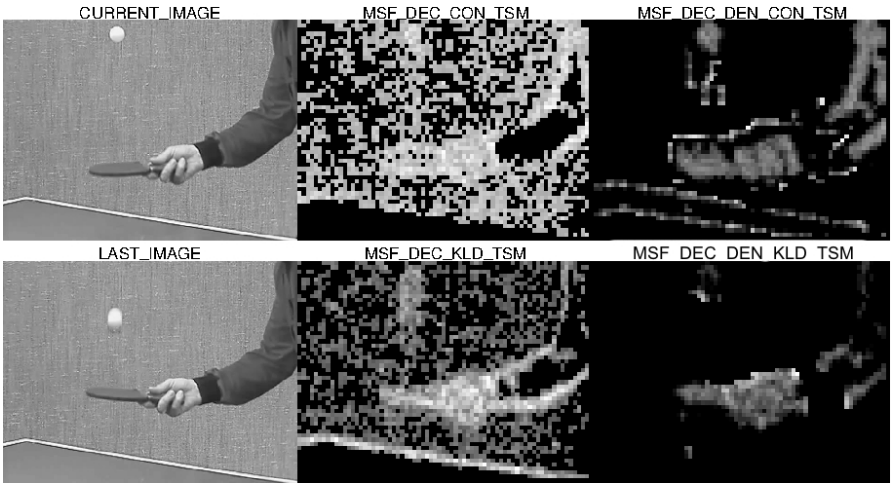


Figure 3.10: Spatial Saliency Map: (left) without denoising, (right) with denoising

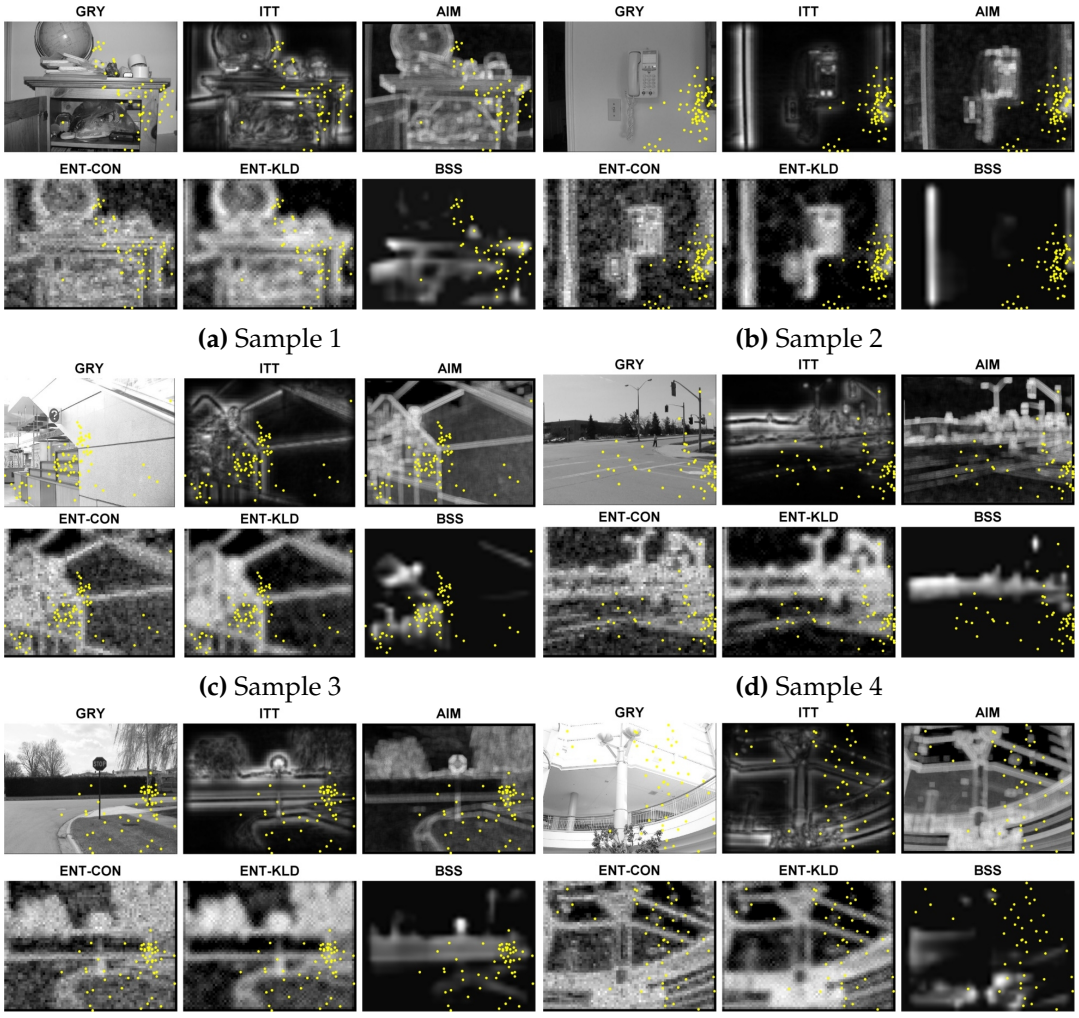
Algorithm 4 Medium Sub-band Frequency Filter - Temporal Data Decorrelation & Denoisation

Require: $input\_frames$
$msf3d\_filter\_decorrelation(input\_frames)$
$wavelet\_coefficients = waveletcdf97(input\_frames, number\_wavelet\_level);$
$wavelet\_coefficients\{L1\} \leftarrow 0;$
$wavelet\_coefficients\{DC\} \leftarrow 0;$
$dct\_coefficients\_reduced = temporal\_decorrelation(wavelet\_coefficients);$
for $dct\_basis\_counter = 1$ **to** $number\_of\_dct\_basis\_reduced$ **do**
$dct\_coefficients\_reduced\_denoised =$
$bivariate\_shrinkage(dct\_coefficients\_reduced, number\_wavelet\_level - 1);$
end for
$msf\_denoised\_feature =$
$inverse\_waveletcdf97(dct\_coefficients\_reduced, number\_wavelet\_level);$
Ensure: $msf\_decorrelated\_feature$

[115]. Evaluation results are compared against well-known saliency methods in the literature including, Itti and Koch (ITT) [1], An Information Maximization method (AIM) [4], Bayesian Surprise Saliency (BSS) [109], MIT saliency method [5], Phase Frequency Transform (PFT) and Phase Quaternion Frequency Saliency (PQFT) [2]. Among these saliency methods, PQFT and the proposed method incorporate both spatial and temporal information meanwhile, others solely generate saliency maps in frame-by-frame manner. Therefore, experiments on both spatial and spatio-temporal data are in need. AIM and MIT are chosen for comparison due to choices of evaluation databases Bruce's database [4] and Judd's database [5]. Moreover, AIM is also an information-based saliency method like the proposed ENTs methods. In addition, BSS, which shares similarities with ENTs in information estimation, is also involved in the comparison.

3.6.1 Experiments on Image Databases

In Bruce's image database [71], eye tracking data were collected from 20 human subjects over 120 different colour images. Scenes of these photos include both outdoor and indoor environment in which some contain very clear and big salient objects, but others are full of various small textures. This image database has been utilized by Bruce [4] for AIM and many other researchers for evaluating their proposed saliency methods.



(a) Sample 1 **(b)** Sample 2

(c) Sample 3 **(d)** Sample 4

(e) Sample 5 **(f)** Sample 6

Figure 3.11: Saliency Maps of Sample Images in Bruce' database [4]

Figure 3.11 shows the saliency maps of five different saliency methods for six sample images of which first three are indoor scenes and last three are out-door scenes. Moreover, all images are shown in gray-scale with overlaid fixation points (yellow points). Visually, ITT, AIM, ENT-CON and ENT-KLD produce quite similar saliency maps while BSS produces saliency maps with less details. Though, saliency maps of ENT-CON and ENT-KLD highlight most details of images, they also include more noises than others.

As visual saliency maps only give subjective judgement of individual saliency maps, it does not demonstrate performance on the whole database of images. Therefore, standard quantitative evaluations LCC, NSS, AUC, TIME and ROC curves on the whole database are carried out to fairly compare saliency methods. The evaluation results are shown in Table 3.2 and plotted in Figure 3.12.

First, ROC curves of saliency maps from different approaches are shown in Figure 3.12. According to this plot, ENT-CON has the best ROC curve and it is confirmed by its highest AUC value in Table 3.2. ENT-CON has high AUC result since its saliency maps highlight many details of scenes. Meanwhile, ROC curve and AUC results of ENT-KLD are worse than ENT-CON since details of saliency maps are blurred. However, both ENT-CON and ENT-KLD produce noises in their saliency maps which affect saliency values at saccade points. Therefore, their NSS results are worse than those of ITT, AIM and BSS methods. In addition, ENT-CON and ENT-KLD are only better than AIM but worse than both ITT and BSS in LCC evaluation, which means their saliency maps linearly correlate with eye-fixation maps better than AIM saliency but worse than those of ITT and BSS. When comparing computational time (TIME) among saliency approaches in the Table 3.2, AIM has the longest processing time duration per frame, roughly 66 seconds, due to computational loads of its information estimation. Meanwhile, ENT with CON and KLD for information extraction only takes roughly 0.9s and 0.5s. In computational time, ENT-CON (0.9s) and ENT-KLD (0.5s) are as efficient as ITT (1.2s) and more efficient than BSS (5.375 s). ENT-CON and ENT-KLD took the least computation time for a sample image averagely among three information-based saliency methods ENT, AIM and BSS. It is noted that experiments of ENT, AIM and ITT are done by MATLAB 2010b in a 2.33 GHz Intel Core 2 Duo computer running Linux Ubuntu 10.10 OS and BSS experiment is done by C++ program in the same computer but with Mandriva OS. ROC curves and AUC values are only useful for comparing different computational saliency approaches, but they don't emphasize comparison between these methods with generalized human performance (correlation across subjects). Therefore, the inter-subject ROC, curves proposed by Harel *et al.*[49], is shown in the Figure 3.12 to compare performance of human visual system versus that of computational saliency method. Human performance is shown in black solid line; meanwhile ISROC curves of ENT-CON, ENT-KLD, AIM and ITT are plotted with different colours. ISROC curves are only above the black line up to their intersected points and below this line after their intersection. The inter-section point between an ISROC curve of each method and the black solid line is where performances of computational saliency methods and humans in predicting eye-fixation points are equivalent. The further that intersection point is toward the top-right corner, the better a saliency method performs than human

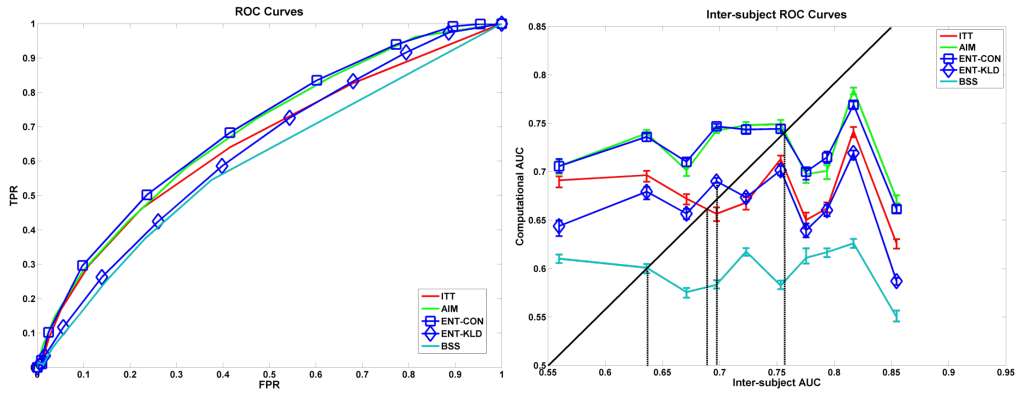


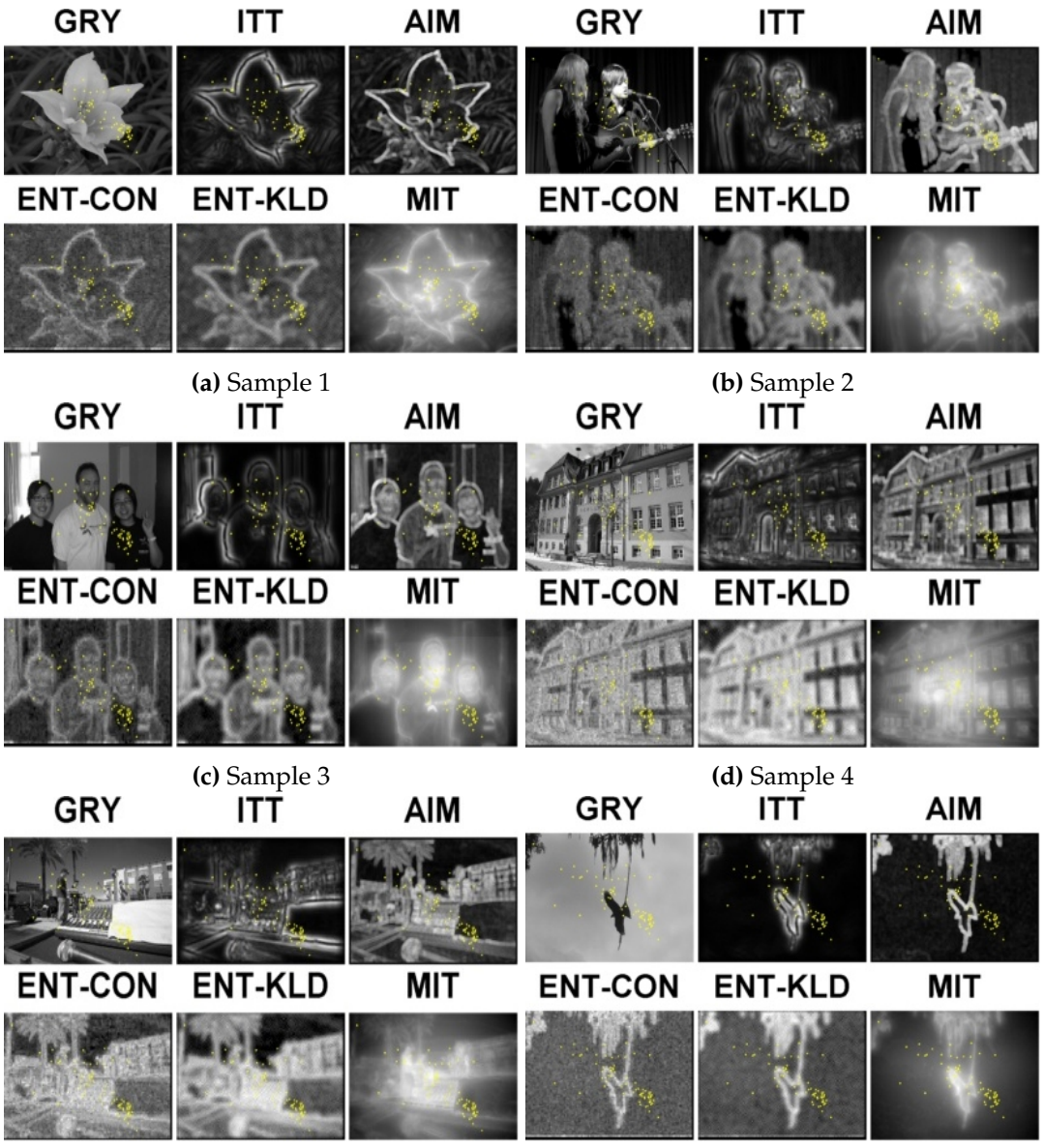
Figure 3.12: ROC and Inter-subject ROC curves on Bruce's database

Table 3.2: Evaluation results of Saliency Methods on Bruce's database

Methods	ITT [1]	AIM [4]	ENT-CON	ENT-KLD	BSS[109]
LCC	0.03492	0.01576	0.02263	0.03071	**0.04823**
NSS	0.27714	0.12378	-0.17533	-0.19000	**0.54031**
AUC	0.6781	0.72391	**0.72238**	0.66486	0.59728
TIME (s)	1.2488	66.2673	0.93094	0.56824	5.37500

beings does. According to the Figure 3.12, ENT-CON and AIM have equivalent performances and only better than human-performances when inter-subject AUC is less than 0.75. Meanwhile, ENT-KLD and ITT have similar performances and only better than human-performances when inter-subject AUC is less than 0.675. It is noted that these mentioned values of inter-subject AUC are found by projection of intersection points on a horizontal of the ISROC plot, as shown in Figure 3.12.

Besides Bruce's database, ENT methods are also evaluated with the Judd's database [5], which has 1003 photos with eye fixation data from 15 viewers. The database was used for evaluating Judd's learning based saliency method [5]. Qualitative (visualization of saliency maps) and Quantitative evaluations (LCC, NSS, AUC and TIME) are carried out on the 100 testing images, reported in Judd's paper [5]. Visual illustration of six gray-scale samples (GRY) , including three indoor scenes and three outdoor scenes, and their saliency maps of ITT, AIM, ENT-CON, ENT-KLD and MIT [5] is shown in Figure 3.13. Eye-fixation points, which are ground-truths of the evaluation, are marked with yellow points, overlaid on sample images and saliency maps. For quantitative evaluations on Judd's database as shown in Table 3.3, AUC results again confirm that the proposed ENT-CON and ENT-KLD methods well compete against other methods in AUC values. However, ENTs are in general perform worse than ITT, AIM and MIT in LCC and AUC evaluations due to its noises. Meanwhile, ENT-CON and ENT-KLD are comparable to ITT in processing time per image; both ENTs are much faster than AIM. It is noted that TIM of MIT method is not reported in the Table 3.3 since MIT saliency maps [5] are provided in the database and not generated.



(a) Sample 1 **(b)** Sample 2

(c) Sample 3 **(d)** Sample 4

(e) Sample 5 **(f)** Sample 6

Figure 3.13: Saliency Maps of Sample Images in Judd's database [5]

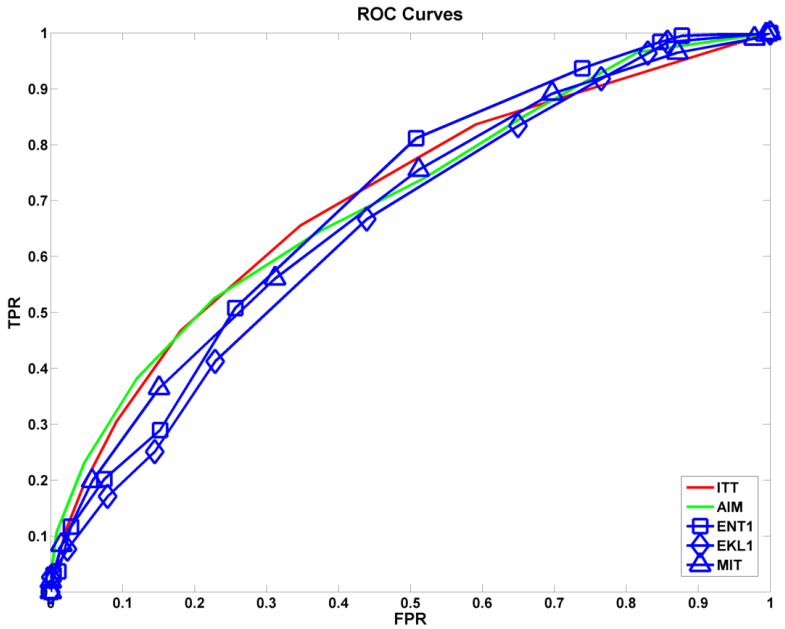


Figure 3.14: ROC curves on Judd's database

Table 3.3: Evaluation results of Saliency Methods on Judd's database [5]

Methods	ITT	AIM	ENT-CON	ENT-KLD	MIT[5]
LCC	0.08097	0.06629	0.05170	0.06179	0.22412
NSS	0.53879	0.30191	0.05877	0.06246	0.94336
AUC	0.74940	0.71165	0.78157	0.69225	0.68845
TIME	0.19138	120.22787	0.20135	0.09735	NaN

3.6.2 Experiments on Video Databases

Aforementioned experiments only focus on spatial aspect of current frames. Generally, past saliency methods were built for still images; it is no longer true and applicable in this era of rapidly escalating video contents which requires faster saliency methods for spatial-temporal information. Therefore, the proposed ENT methods are extended to estimate saliency temporally in Section 3.4. For quantitative performance evaluations and comparisons of relevant saliency methods, simulations must be done only for spatio-temporal data, video materials. Moreover, this work focuses on specific video context, human visual perception in driving contexts; therefore, the supplementary video data must be recorded in drivers' point of views and psychological related data eye-fixation or ground truth segments done by human beings. Though there are several datasets for saliency evaluation from Bruce *et al.*[4], Judd *et al.*[5], Koostra *et al.*[23], etc, most of available datasets are still images which presented to subjects over a period time. Then subjects are asked to do a free-viewing task while their eye-fixations are recorded. Therefore, most of standard datasets [116] are not suit-

able for evaluating the spatio-temporal ENT saliency method in driving scenes. Fortunately, there are two proper databases AUTOM and MSDR [115] for this evaluation task. AUTOM database includes approximately 10 minutes videos, recorded by a camera on head-mounted eye-trackers which simultaneously detect drivers' fixation points on road scenes. Meanwhile, MSDR is as well recorded on road with drivers' view in high resolution but it has ground truth for each frame instead of eye-fixation data. Experiments on AUTOM database emphasize correlation between artificial saliency maps of different computational method with human psychological data; meanwhile, MSDR with ground truths, which are segmented by volunteers, stresses on specific correlation between saliency maps and important features of the scenes.

AUTOM database

In AUTOM database, the video contents are created by Accident Research Unit, the University of Nottingham, UK Campus, which later cut into 28 short videos. Each has 500 frames except the last video which has 512 frames. The whole video database has 14012 frames, lasting for 9 minutes and 20 seconds. These videos are recorded with a head-mounted camera of SMI eye-trackers; simultaneously, two small cameras, pointing to pupils of the driver's eyes, locate drivers' eye fixation at 250 Hz. The videos are recorded in real-life driving situations; therefore, several road types and traffic situations have been covered such as urban roads, high ways, speeding up and slowing down situations. It does not extensively cover all possible driving schemes, but provides sufficient materials about human eye-fixation in most daily driving situations. Importantly, each frame of these videos is recorded with drivers' eye fixation response to specific real-life driving situations. Noteworthy, different from collecting multiple eye-fixation data for each static picture, only one fixation point is collected for each frame. The data format is changed, so are evaluation methods since ROC, AUC and ISROC are not entirely suitable for the evaluation of video with eye-marks.

In saliency literature, Normalized Saliency Value (NSV) and Chance Adjusted Saliency (CAS) [117] have been proposed for correlating human performance and generated saliency maps spatio-temporally. Normalized Saliency Value is collecting saliency values of normalized saliency maps at locations of eye-fixations; the larger NSV is, the more accurate saliency methods can predict human attention points. However, if the saliency value is directly retrieved at locations of eye-marks, they are prone to a mean spatial error of eye-trackers usually $0.5^\circ - 1^\circ$. In order to eliminate those errors, the maximum saliency value in square areas, whereof a fixation point is at the centre, is extracted. In Table 3.4, a row $Sample_{\#\#}$ represents normalized saliency values of the

Table 3.4: Normalized Saliency Value

Observations	ITTI	GBVS	PFT	PQFT	ENT-CON	ENT-KLD
sample_00	0.026	0.39	0.145	0.146	0.398	0.596
sample_01	0.018	0.339	0.131	0.136	0.318	0.444
sample_02	0.033	0.359	0.156	0.156	0.489	0.671
sample_03	0.028	0.406	0.185	0.195	0.556	0.708
sample_04	0.01	0.306	0.099	0.100	0.242	0.348
sample_05	0.029	0.280	0.080	***0.088***	0.209	0.271
sample_06	0.028	0.290	0.096	0.122	0.254	0.364
sample_07	0.008	0.371	0.125	0.164	0.267	0.390
sample_08	0.022	0.378	0.141	0.161	0.323	0.468
sample_09	0.014	0.433	0.156	0.169	0.411	0.548
sample_10	***0.002***	0.306	0.077	0.096	***0.145***	***0.240***
sample_11	0.015	0.336	0.107	0.101	0.241	0.341
sample_12	0.034	0.387	0.139	0.13	0.369	0.545
sample_13	0.015	0.380	0.142	0.134	0.338	0.435
sample_14	0.017	0.381	0.215	0.163	**0.564**	0.690
sample_15	0.064	0.409	0.174	0.185	0.563	0.704
sample_16	0.024	***0.257***	0.082	0.101	0.216	0.353
sample_17	0.027	**0.527**	0.231	**0.271**	0.555	0.736
sample_18	0.022	0.423	0.157	0.198	0.327	0.442
sample_19	0.065	0.286	***0.060***	0.103	0.159	***0.240***
sample_20	0.123	0.423	0.123	0.175	0.363	0.582
sample_21	**0.160**	0.374	0.118	0.151	0.36	0.610
sample_22	0.059	0.341	0.099	0.157	0.282	0.456
sample_23	0.077	0.354	0.135	0.155	0.355	0.643
sample_24	0.048	0.411	0.177	0.189	0.409	0.589
sample_25	0.069	0.389	0.194	0.192	0.475	0.695
sample_26	0.019	0.415	0.238	0.239	0.482	0.628
sample_27	0.097	0.418	**0.243**	0.244	0.558	**0.746**
Mean	0.041	0.370	0.144	0.158	0.365	0.517
Variance	0.014	0.033	0.025	0.0021	0.016	0.025

##th video in the AUTOM database with respect to six different saliency approaches. The NSV values in each video are averages of those across all frames. The last two rows show mean and variance of saliency values at eye-fixations in consideration of the whole 10 minutes video. Generally, either ENT-CON or ENT-KLD saliency methods outperform the other state-of-the-art approaches. Second to the proposed method in average NVS values is Graph-based Visual Saliency (GBVS) method [49] , then PQFT, PFT and ITT saliency methods are ranked in descending orders.Moreover, **maximum** and ***minimum*** NSV values are formatted accordingly for each column of the Table 3.4. The best and worst performances of each method across all video samples highlight consistency in performance between ENT-CON / ENT-KLD and the others. For example, ITT, ENT-CON and ENT-KLD has the worst NSV values at sample 10 among all samples, and NSV values of GBVS (0.306), PFT (0.077), and PQFT (0.096) in this sample are also far lower than their respective mean values (GBVS: 0.37 $\pm$ 0.033; PFT: 0.144 $\pm$ 0.025; PQFT: 0.158 $\pm$ 0.0021). Meanwhile, ENT-KLD and PFT has the best NSV values at sample 27 among all samples, and NSV values of ITTI (0.097), GBVS (0.418), PQFT (0.244), and ENT-CON (0.558) are far higher than their respective mean values (ITTI: 0.041 $\pm$ 0.014; GBVS: 0.37 $\pm$ 0.033; PQFT: 0.158 $\pm$ 0.0021; ENT-CON: 0.365 $\pm$ 0.016).

Chance-adjusted saliency (CAS) performance metric, proposed by Parkhurst *et al.*[50], emerges in literature as the preferred metric for saliency method comparisons. It basically measures difference in correlation of each saliency method between fixation points generated by human visual systems and random number generators. The more significant the difference (CAS) is, the better saliency maps can distinguish between human eye fixations and randomly generated fixations. According to means and variances in the last two rows of Table 3.5, GBVS saliency method has more success than other saliency methods in distinguishing between human and random eye-fixations. Second to GBVS are the proposed methods ENT-CON and ENT-KLD, then ITTI, PFT and PQFT are ranked in descending order.In addition, **maximum** and ***minimum*** CAS value of each column in Table 3.5 are formatted accordingly. The best and worst CAS value shows consistency in performances between proposed saliency methods and the others. For example, ITTI, PFT, PQFT, ENT-CON, ENT-KLD has the least CAS value at sample 10 while GBVS's CAS value (0.08053) is far less than its mean (0.14905 $\pm$ 0.00314). Meanwhile, PFT, ENT-CON,ENT-KLD have their best CAS values at sample 27 and CAS values of ITTI (0.05362), GBVS (0.21951) and PQFT (0.05795) are much larger than their respective means (ITTI: -0.00170 $\pm$ 0.00136; GBVS: 0.14905 $\pm$ 0.00314;

Table 3.5: Chance Adjusted Saliency

Observations	ITTI	GBVS	PFT	PQFT	ENT-CON	ENT-KLD
sample_00	-0.01620	0.15105	-0.01912	-0.02068	0.09810	0.16031
sample_01	-0.02883	0.10282	-0.05274	-0.05878	-0.00154	-0.01592
sample_02	-0.00547	0.13573	-0.03199	-0.03077	0.16424	0.20673
sample_03	-0.01498	0.19607	0.03642	0.02633	0.27253	0.29317
sample_04	-0.03208	0.08501	-0.08824	-0.08548	-0.07596	-0.10103
sample_05	-0.01848	0.05986	-0.11093	-0.08703	-0.09415	-0.15169
sample_06	-0.02129	0.07737	-0.05542	-0.04822	-0.03347	-0.04137
sample_07	-0.02886	0.12561	-0.04424	-0.04983	-0.06474	-0.06572
sample_08	-0.01703	0.13734	-0.04811	-0.04805	-0.01329	0.01062
sample_09	-0.02898	0.18808	-0.03441	-0.04354	0.04525	0.06153
sample_10	***-0.04026***	0.08053	***-0.11124***	***-0.12110***	***-0.22137***	***-0.26291***
sample_11	-0.02833	0.09366	-0.09500	-0.08694	-0.07029	-0.10209
sample_12	-0.00883	0.16270	-0.03029	-0.03776	0.07808	0.12318
sample_13	-0.02912	0.16786	-0.02083	-0.02337	0.04448	0.02625
sample_14	-0.02420	0.15050	0.01061	-0.02473	0.22266	0.22460
sample_15	0.01803	0.17338	-0.00786	0.00165	0.24206	0.26677
sample_16	-0.01952	***0.05000***	-0.08519	-0.07502	-0.08264	-0.06527
sample_17	-0.01150	**0.27638**	0.03338	**0.09682**	0.25085	0.28031
sample_18	-0.02229	0.16684	-0.02129	0.00881	0.03661	0.00775
sample_19	0.01972	0.07822	-0.08941	-0.07300	-0.10352	-0.16421
sample_20	0.08400	0.21136	-0.01114	0.00762	0.10156	0.16437
sample_21	**0.11576**	0.20080	-0.00166	0.00948	0.12431	0.24193
sample_22	0.01808	0.13805	-0.01966	-0.01490	0.02612	0.06633
sample_23	0.03123	0.14159	-0.01558	-0.02963	0.05672	0.20264
sample_24	0.00010	0.22466	0.03073	0.01936	0.14511	0.18615
sample_25	0.02951	0.19172	0.03718	0.03137	0.21777	0.30256
sample_26	-0.02134	0.18658	0.07070	0.05061	0.18509	0.19666
sample_27	0.05362	0.21951	**0.09223**	0.05795	**0.28961**	**0.34264**
Mean	-0.00170	0.14905	-0.02440	-0.02317	0.06572	0.08551
Variance	0.00136	0.00314	0.00273	0.00252	0.01779	0.02716

PQFT: -0.02317 $\pm$ 0.00252).

$$R_{NSV} = \{ENT-KLD, ENT-CON, GBVS, PQFT, PFT, ITTI\}$$
$$R_{CAS} = \{GBVS, ENT-KLD, ENT-CON, ITTI, PQFT, PFT\}$$

By quantitative observations of mean values and variance, it is concluded that that groups of saliency methods GBVS, ENT-KLD and ENT-CON outperforms the other group PQFT,PFT, and ITTI. Moreover, there is reasonable trend inside each group when CAS or NSV are chosen as evaluating methods. For example, the proposed method generates saliency maps with more specific local details than GBVS does; therefore, ENT-KLD / ENT-CON performs better than GBVS in NSV but worse than GBVS in CAS. Similarly, PFT and PQFT are superior to ITT in NSV while they perform worse that ITT in CAS. In order to make the comparison fairer between local feature saliency methods and their global counterparts, region of interest mask are applied on local saliency maps. The ROI masks are simply generated by setting threshold on saliency values. Similar explanation can be used for discrepancy in CAS, and NSV results of inferior group of saliency methods.

Table 3.6: Normalized Saliency Values in various driving contexts

Observations	ITTI	GBVS	PFT	PQFT	ENT-CON	ENT-KLD
go_alone	0.01851	0.33936	0.14633	0.13281	0.37296	0.49269
head_toward	0.02014	0.38734	0.15944	0.16283	0.45549	0.60256
go_away	0.06923	0.36038	0.11404	0.14979	0.28620	0.45065
overtake	0.07783	0.43606	0.14036	0.19497	0.43177	0.59839
slow_down	0.02451	0.41908	0.16767	0.16924	0.45894	0.59269
follow_car	0.03816	0.37397	0.16606	0.16504	0.36305	0.52260
Mean	0.04140	0.38603	0.14898	0.16245	0.39473	0.54326
Variance	0.00056	0.00110	0.00034	0.00036	0.00374	0.00343

Table 3.7: Difference in Normalized Saliency Values between go_alone and others

Observations	ITTI	GBVS	PFT	PQFT	ENT-CON	ENT-KLD
go_alone	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
head_toward	0.00146	0.04781	0.01294	0.02982	0.01619	0.02730
go_away	0.04952	0.02009	-0.03281	0.01635	-0.01624	-0.00526
overtake	0.04721	0.04149	0.01184	0.03407	0.00109	0.05625
slow_down	0.00996	0.02267	0.01569	0.04101	0.01768	0.05115
follow_car	0.01574	0.04859	0.00204	0.02750	-0.00926	-0.01375
Mean	0.02065	0.03011	0.00162	0.02479	0.00158	0.01928
Variance	0.00041	0.00031	0.00027	0.00018	0.00015	0.00075



(a) go_alone **(b)** follow_car

(c) slow_down **(d)** go_toward

(e) go_away **(f)** overtake

Figure 3.15: Attention maps generate by six different methods

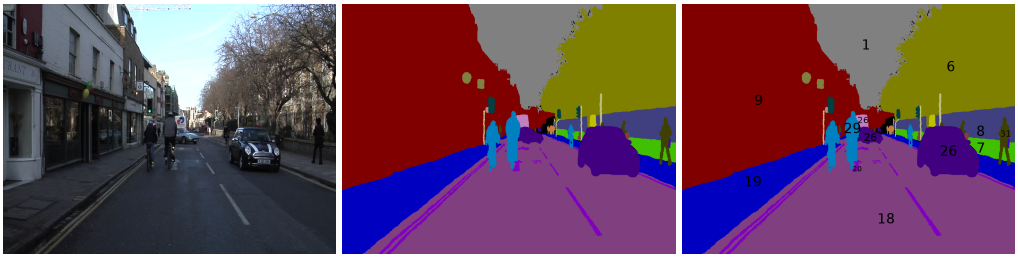
Besides the studies in general driving scenes, statistical correlation between saliency maps and human responses to specific driving situations are interested as well. It helps answer the question how specific driving situations may affect relations between saliency maps and human eye fixations. Does that correlation get better or worse specially? Before answering this question, available driving contexts in the recorded data needs to be specified, and the number of frames for each case should be large enough to ensure the relative generality of these experiments. After skimming throughout the database, there are six significant specific driving situations: `go_alone`, `go_toward`, `go_away`, `slow_down`, `overtake` and `follow_car`.
These situations are named after main tasks which drivers carry out; for example, `go_alone`, `slow_down` and `follow_car` are used to describe situations wherein a car is driven alone without any other moving objects on roads - the figure 3.15a, is slowed down when reaching a street corner - the figure 3.15c, and the driver follows a car in front - the figure 3.15b. Likewise, `go_toward` or `go_away` are used for in-front cars which are moving toward the driver 's car - the figure 3.15d, or moving away the driver's car - the figure 3.15e. Similarly, `overtake` is named after the situation wherein a behind car suddenly overtakes the driver's car - the figure 3.15f. In the `go_alone` context, there are no other moving objects on roads which means less visually salient objects and values at eye-marks of those scenes. In other words, the normalized saliency values of *go_alone* should be less than any other cases. Table 3.7 summarizes differences in normalized saliency values (NSV) at human fixation points, Table 3.6, between `go_alone` and other driving situations. Almost all values in Table 3.7 is positive increment in performance of saliency methods whenever drivers are found in particular traffic situations rather than being alone on the roads. The ENT-KLD and ENT-CON saliency method still possesses the best normalized saliency values in Table 3.6, and its performance in specific driving contexts are generally better than `go_alone` situations. The decrements of evaluation results in some specific situations may be due to limitation of frames in that context. More video samples need to be collected and analysed later to clearly find out how saliency maps and human eye-fixations are related in a particular situation. In the constrained number of input video samples and experiments, saliency methods and human psychological data generally have more matches in specific driving situations than being alone on the roads.

MSRD Database

In the previous AUTOM database section, studies and judgements mainly based on psychological fitness of computational saliency maps according to availability of eye-



Figure 3.16: Samples and Saliency Maps of ITT, GBVS, PFT, and PQFT, and ENT methods (left-to-right top-to-bottom order). Please see the supplementary materials for sample sections of the video containing the original scene with the eye fixation marker overlayed on it and saliency maps of various methods.



(a) Sample Image **(b)** Ground Truth **(c)** Importance Map

Figure 3.17: MSDR Data Samle Image, Ground Truth, Importance Map

fixation data. In this section, a study, focusing on application aspect of saliency methods, is done on Motion-based Segmentation and Recognition dataset (MSRD) [118] with their ground-truth data; figure 3.17 reveals a sample of a video from MSDR database with its ground truth.

Before going into specific details of its evaluation, it is necessary to briefly describe the MSRD data as well as rationales of this experiment. Another name of MSRD database is Cambridge-driving Labeled Video Database (CamVid), a collection of driving context videos with semantic metadata of object classes. It provides ground-truth labels that associate each pixel with one of 32 semantic classes. Contrary to majority of traffic database recorded as surveillance footage from CCTV camera, it uses similar approach to aforementioned AUTOM database by capturing video data from the perspective of a driver. Comparing to AUTOM, its driving scenarios are increasingly complex and filled with many objects in daily driving scenes. Its heterogeneity is suitable for studying saliency maps in general driving scenarios. Furthermore, MSRD database provides over ten minutes of high quality 30 Hz footage wherein some parts are semantically

labelled at 1Hz and 15Hz. Totally, ground truth data of over 700 images are manually specified by the first human observer and reconfirmed by the second subject for accuracy. Though videos of the whole database last for ten minutes, only 700 sample frames with their correspondent ground truth data are useful in these studies. In addition to 700 frames extracted from original video at 1Hz or 15Hz, four of their past frames are acquired to serve as input data for spatio-temporal saliency maps.

In MSRD dataset, ground truth data of each sample are results of manually classifying objects by volunteers. In ground-truth maps, each pixel belongs to one of 32 available classes. For a specific application, a certain class of objects will be more important than others; for instance, pedestrians class rather than other classes is vital for pedestrian detection. Therefore, these studies of general driving contexts and safety implies a certain importance values for 32 available classes. Among those, there are roughly four categories: *a group of moving objects* , .i.e pedestrian, children, *a group of road objects* , .i.e road, road-shoulder, *a group of ceiling objects*, sky, tunnel, and *a group of fixed objects*, building, wall, tree. Generally in driving situations, those groups can be arranged as moving objects, road objects, fixed objects, and ceiling objects in descending order of their importance. Inside each group, object classes are ranked for its vitality with respect to driving tasks in the scene; then, there are importance scores for every available object class in Table 3.8. Using Table 3.8, image pixels of specific classes in the ground

Table 3.8: Importance order of MSRD object classes

Moving objects	**Imp**	**Road objects**	**Imp**	**Fixed objects**	**Imp**
Child	32	Road	18	TrafficLight	17
Pedestrian	31	RoadShoulder	19	SignSymbol	16
Animal	30	LaneMkgsDriv	20	TrafficCone	15
Bicyclist	29	LaneMkgsNonDriv	21	Column_Pole	14
MotorcycleScooter	28	**Ceiling objects**	**Imp**	Sidewalk	13
CartLuggagePram	27	Archway	3	Bridge	12
Car	26	Tunnel	2	ParkingBlock	11
SUVPickupTruck	25	Sky	1	Misc_Text	10
Truck_Bus	24			Building	9
Train	23			Fence	8
Misc	22			Wall	7
				Tree	6
				VegetationMisc	5
				Void	4

truth data can be mapped to its corresponding importance values. In other words, an Importance Map (IM) can be constructed from ground truth data of MSRD dataset. Currently, each pixel in IMs has integer values in the range of $[1, 32]$, while saliency

values are decimal numbers in the range $[0, 1]$. Then values of IMs need scaling down to $[0, 1]$ to ensure consistency of further evaluations.

Normalized importance maps show us how relatively important each pixel on that map is for driving context. Moreover, it is necessary to evaluate whether or not saliency values are proportional to importance degree of ground-truth maps for arbitrary regions. In other words, it is mathematical problem of identifying how well correlated importance and saliency maps are. The normalized cross correlation is chosen to measure and demonstrate that relation. Assumed that there is a saliency map S in short of $S(x, y)$ and importance map M in short of $M(x, y)$, the normalized cross correlation (NORMXCORR) is formulated as the Equation 3.6.1, and its range of value is $[-1, 1]$.

$$\text{NORMXCORR} = \frac{\sum_{x,y}[S - \mu(S)][M - \mu(M)]}{(\sum_{x,y}[S - \mu(S)]^2 \sum_{x,y}[M - \mu(M)]^2)^{\frac{1}{2}}} \tag{3.6.1}$$

Among over 700 useful data images, they belong to three video scenes. In order to visualize the simulation results, NORMXCORR of frames from three video sequences will be plotted in three Figures 3.18a,3.18b,3.18c. Each figure has six lines with different colours representing NORMXCORR between importance and saliency maps from six saliency methods. These plots show how the measurement fluctuate and change across the whole video, and their statistical mean and variance are also displayed in the Tables 3.9, 3.10, 3.11. After looking through three plots and tables, NORMXCORR

Table 3.9: Statistical data of Normalized Cross Correlation over video sequence 01

Observations	ITTI	GBVS	PFT	PQFT	ENT-KLD	ENT-CON
Mean	-0.09315	-0.16580	-0.26443	-0.17535	-0.09301	-0.10431
Variance	0.00615	0.00775	0.00620	0.02255	0.00188	0.00248

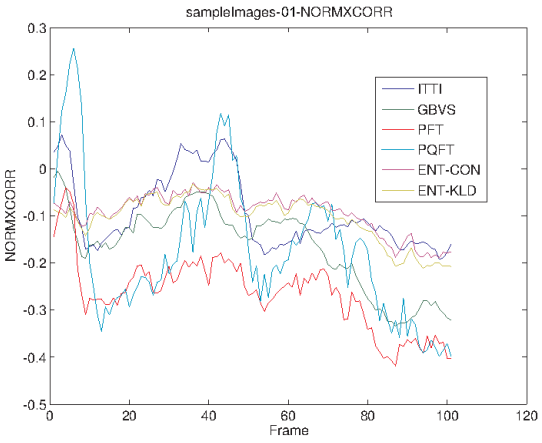
Table 3.10: Statistical data of Normalized Cross Correlation over video sequence 02

Observations	ITTI	GBVS	PFT	PQFT	ENT-KLD	ENT-CON
Mean	0.02512	-0.08502	-0.03021	0.02062	0.20900	0.21499
Variance	0.00866	0.02305	0.04298	0.04906	0.04088	0.04544

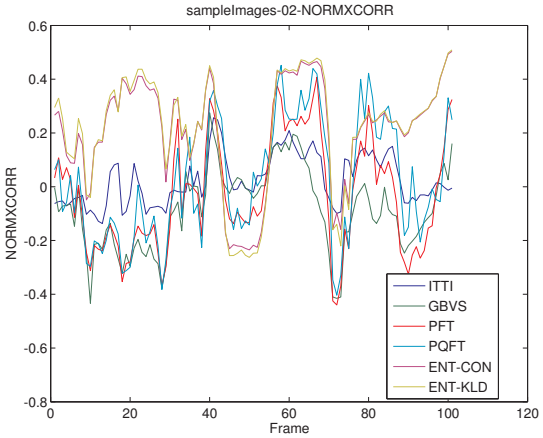
Table 3.11: Statistical data of Normalized Cross Correlation over video sequence 03

Observations	ITTI	GBVS	PFT	PQFT	INFO	ENTRO
Mean	-0.04801	-0.26006	-0.30763	-0.23102	0.24872	0.25500
Variance	0.00638	0.00763	0.01083	0.01970	0.02723	0.02761

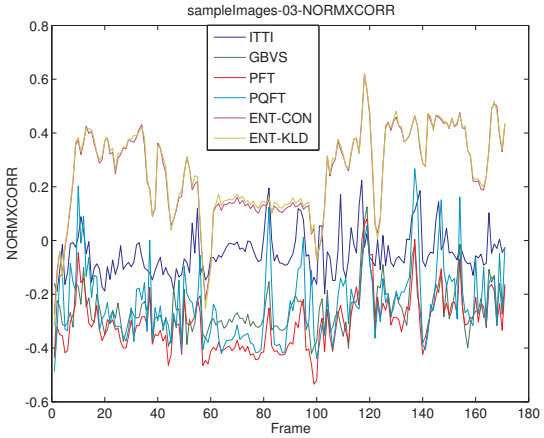
of proposed methods is larger than the other established methods. Though in some psychological measurements, the proposed ENT method does not give the best performance, its saliency maps are well correlated with application oriented importance maps. In other words, it provides useful clues about how important and meaningful



(a) Sequence_01



(b) Sequence_02



(c) Sequence_03

Figure 3.18: Normalized Cross Correlation of Importance Map and Saliency Map

each pixel is to the scene.

3.7 CONCLUSION

In brief, this chapter presents interpretation of bottom-up visual saliency as entropy-based saliency methods (ENT) in which centre-surround operator is formulated as conditional entropy (CON) and Kullback-Leibler divergence (KLD). Moreover, it also suggests a fast information estimation method for computing saliency maps of images and videos despite of their high-dimensionality. The curse of dimensionality causes the estimation method sensitive to noisy data; therefore, several pre-processing techniques are utilised to partly remove noises from input data. Experiments show that the proposed methods (ENT-KLD & ENT-CONT) are efficient in prediction of eye-fixation points, as shown by high AUC values and good ROC and ISROC curves. However, saliency maps of ENT-CON and ENT-KLD are still affected by generated noises from

information estimation methods; which in turn worsen their performances in LCC and NSS values. Moreover, the ENT techniques are feasibly extended for video data, spatial-temporal saliency and their performances are evaluated on two driving context video database AUTOM and MSRD. These experimental results show that ENT-KLD and ENT-CON saliency maps correspond well to eye-fixations of on-board drivers.

CHAPTER 4

Computational Model for Wavelet-based Scale Saliency (WSS)

Pixel-based scale saliency (PSS) models [22] saliency measurement from both image features contained in a window and their variations across multiple scales of that window. In PSS, image features are the original pixel values; however, other basis-projected features such as wavelet coefficients are eligible for computing the scale saliency as well. So far, no models have ever been proposed for calculating scale saliency on wavelet coefficients, basis-projected descriptors. Therefore, This chapter proposes extension of those ideas into mathematical models and implement them in the wavelet-based scale saliency (WSS) . While PSS uses pixel-value descriptors, WSS treats wavelet energy sub-bands as basis descriptors which include different wavelet transform: discrete wavelet transform (DWT) , discrete wavelet packet transform (DWPT) , quaternion wavelet transform (QWT) and best basis quaternion wavelet packet transform (QWPTBB) . WSS saliency maps of different descriptors are generated and compared against other saliency methods by both quantitative and qualitative methods. Quantitative results, ROC curves, AUC values and NSS values are collected from simulations on Bruce and Kootstra image databases [18] [119] with human eye-tracking data as ground-truth. Furthermore, qualitative visual results of saliency maps are analysed and compared against each other as well as eye-tracking data inclusive in the databases.

4.1 INTRODUCTION

Information-based saliency approaches all motivated from the assumption that human attention could be attracted to spatial location accompanied with highly informative

content. From signal coding, compression and self-information theory, an event has more information when it appears to be structural-persistent across multiple scales and rarely appear anywhere else on an image domain. Pixel-based Scale Saliency (PSS) [22] is so far the only approach accumulating information of both descriptors and their persistent structure across scale-spaces. However, PSS employs pixel-value descriptors and isotropic circular sampling, which might hinder its performance in term of accuracy due to its failure in extracting popular oriented features in natural images as well as speed due to the curse of dimensionality in information estimation.

The limitation of PSS has sparked discussion for its extension by Kadir *et al.*[22] [53]. It is suggested that basis-projection method can provide better descriptors for scale saliency techniques. Furthermore, utilization of alternative descriptors could bring in another theoretical viewpoint of scale saliency and multi-scale structural information. Moreover, such extension would make scale saliency the first-ever method capable of using both pixel-value descriptors (PSS) and basis-projected descriptors. This chapter promotes wavelet analysis as a potential basis-projection method and wavelet-based descriptors for Scale Saliency. Therefore, one of the proposed methods is named as Wavelet Scale Saliency (WSS). Besides WSS, a derivative of WSS is also proposed and named as Discriminant Saliency (DIS). In order to clearly explain extension from pixel-based descriptors to wavelet-based descriptor, this chapter is organized in following sections. Section 4.2 gives overview about scale saliency and its main characteristics; meanwhile, Section 4.3 describes details of wavelet-based descriptors and WSS modelling. In Section 4.3, Sub-Section 4.3.1 explains the rationale and advantages of using wavelet-based descriptors in time-frequency domain instead of pixel-based descriptors in time-domain for modelling visual saliency. Furthermore, Sub-sections 4.3.2, 4.3.4 elaborate statistical distribution and correlation of these wavelet-based descriptors. As wavelet analysis is chosen as basis-projection method, Sub-section 4.3.3 gives background information about four types of wavelet transforms considered in this study: discrete wavelet transform (DWT), discrete best basis wavelet packet transform (DWPTBB) as well as two quaternion wavelet transforms (QWT) and quaternion wavelet packet transform with best basis (QWPTBB). Accordingly, there are four different time-frequency descriptors: DWT, QWT, DWPTBB, and QWT. Each descriptor depends on a particular morphological shape of its own mother wavelet , whose properties are further analysed in Section 4.3.4 . Along with these proposed descriptors, suitable mathematical models of feature-space and inter-scale saliency estimation are derived in Sub-Section 4.3.5 which unveils strong theoretical relations between WSS and Bayesian Surprise Saliency (BSS) [93]. Besides theoretical evaluation, simulations on Neil Bruce image database [71] and Kootstra's image database [119] are carried out

in order to compare quantitatively the proposed WSSs with different wavelet-based descriptors against the original PSS, the AIM [4] and the BSS model . Furthermore, qualitative analysis on particular images provides better details about responses of the proposed methods with different types of scenes, and all results with discussion are detailed in the Section 4.4. Finally, the Section 4.5 summarizes the main contributions and future research directions.

4.2 SCALE SALIENCY CONCEPT

To better understand scale saliency; a few fundamental principles of original scales saliency need to be reviewed. Scale saliency ,(Y_D),utilizes maximum feature-space entropy weighted by its inter-scale dependency across scales as saliency values; furthermore, it argues that information measurement might be data-driven pivot for human visual attention. Its mathematical model is summarized as follows.

$$Y_D\left(s_p,\vec{x}\right) = H_D\left(s_p,\vec{x}\right) W_D\left(s_p,\vec{x}\right) \tag{4.2.1}$$

$$H_D\left(s_p,\vec{x}\right) = -\int_{d\in D} p\left(d,s_p,\vec{x}\right)\log_2\left(p(d,s_p,\vec{x})\right)\delta d \tag{4.2.2}$$

$$W_D\left(s_p,\vec{x}\right) = s\int_{d\in D}\left|\frac{\delta p\left(d,s_p,\vec{x}\right)}{\delta s}\right| dd \tag{4.2.3}$$

$$s_p = \{s|\frac{\delta H_D\left(s,\vec{x}\right)}{\delta s} = 0; \frac{\delta^2 H_D\left(s,\vec{x}\right)}{\delta s^2} < 0\} \tag{4.2.4}$$

$Y_D(s_p,\vec{x})$, in Equation 4.2.1, represents scale-saliency value at a specific scale or sampling window size (s) for each image location $(\vec{x})$; meanwhile, feature-space saliency, (H_D) in Equation 4.2.2, is measured by its Shannon entropy of pixel-values descriptor (d). Shannon entropy is chosen since it satisfies four over five criteria of multi-scale entropy filtering [120].

1. Flat signal has zero information.
2. The amount of information in a signal is independent of the background.
3. The amount of information in a signal is dependent on the noise.
4. The entropy must work in the same way regardless of a pixel value.
5. The amount of information is dependent on the correlation in the signal.

The last criterion actually requires structural correlation from information estimation, which is apparently not considered in Shannon entropy. Inter-scale saliency, (W_D), actually fulfils this requirement, and it is estimated at every location by total variation

of descriptors' probability distribution function (PDF) across scales. Then, the scale (s_p) at which most significant information should be found; it is actually the maximum point of the scale-entropy concave curve in the equation 4.2.4. Finally, the overall saliency is stated mathematically as Equation 4.2.1 in accordance with the definition of scale saliency. Let's apply the concept of scale saliency on a general form of signal $R(x_0, s_i) = \{I(\vec{x_0}, s_i) + N(\vec{x_0}, s_i)|i = 1...n\}$, where $I_{\vec{x_0},s_i}$ is ideal noise free signal, $R_{\vec{x_0},s_i}$ is the measured signal with noise $N_{\vec{x_0},s_i}$ at specific location and scale $(\vec{x_0}, s_i)$.Assuming no dependencies between noise and the ideal signal, the estimated entropy is $H_D\left(R_{\vec{x_0},s_i}\right) = H_D\left(I_{\vec{x_0},s_i}\right) + H_D\left(N_{\vec{x_0},s_i}\right)$. Assuming that PDF of noise is scale-invariant, Equation 4.2.3 implies that inter-scale saliency measure is purely dependent on variation of useful signal $\Delta_{s_i} H_D\left(I_{\vec{x_0},s_i}\right)$ and not affected by variation of noise due to $\Delta_{s_i} H_D\left(N_{\vec{x_0},s_i}\right) = 0$. This briefly explains basic motivation behind scale saliency work; further mathematical analysis and experiments results can be found in [22] [53].

The original scale saliency [22] uses pixel-value descriptors which are simple, intuitive, and straight forward interpretation of image data. Moreover, its combination with circular sampling window provides isotropic information analysis, independent of any morphological shape inside sampled regions. Nevertheless, its drawbacks are susceptibility to noise, high computational cost and significant bias in entropy estimation toward isotropic features [53]. So far, histogram construction and approximated Parzel kernel are two popular parameter methods for constructing pixel-value descriptors' PDF and estimating entropy. Entropy bias and speed performance in those mentioned methods greatly depend on manual tuning of histogram numbers of bins or Parzel size kernel; in addition, they as well restrict extension of scale saliency to higher dimensional data. Suau [121] overcomes these problems by bypassing PDF construction stage and estimating PSS by multivariate-data-adaptive technique of information estimation [21]. In spite of its fast computation for multivariate data, the non-pdf approach hinders the inter-scale saliency process which directly depends on construction of PDFs in the Equation 4.2.3. A possible solution is adapting set-theory based elegant solutions of Kadir [53] for inter-scale saliency W_D computation into kd-tree structure. However, it is not yet intuitively and mathematically related to the information-based frame-work. It motivates this proposal of WSS method, a more coherent information-based scale saliency with sub-band energy descriptors, as solutions for all drawbacks of PSS.

4.3 WAVLET-BASED SCALE SALIENCY MODEL

4.3.1 Saliency in Scale Space and Spectral Space

A well-known computational model of visual attention is first mentioned in Koch and Ullman's publication [35]. After that, several other models [122],[123],[38] have been proposed; however, they are usually too complex and not biologically plausible. The disadvantages might be due to pixel representation utilized in many early visual attention algorithm. To overcome these problems systematically, Urban [20] has investigated strong constraints to keep computational complexity within an acceptable range for possible real-time implementation. These constrains are drawn from evidences of psychological experiments which show that images could be analysed in psycho-visual channels at least in TV-viewing condition [124]. In other words, visual data could be further analysed into channels and sub-bands instead of raw pixel format. Furthermore, the channels can be effectively characterized by separated frequency bands and orientation ranges of wavelet analysis [125]. Let's assume visual active areas of brains can deploy Cohen-Daubechies-Feauveau (CDF 9/7) wavelet transform operators; then, it results in multi-scale pyramid composed of oriented contrast maps with limited frequency ranges and low-resolution images. For each level of wavelet decomposition, there are four channels: (i) sub-band 0 is approximated image after filtered with many low-pass blurring kernels; (ii) sub-band 1 extracts horizontal frequencies corresponding to vertical edges of images (iii) sub-band 2 contains frequencies and features along two diagonals of image frames. (iv) sub-band 3 prefers vertical frequencies mapping to horizontal features form images. Natural scenes are full of horizontal, vertical or two diagonals features; therefore, human visual perception seems to prefer those dominant features. Besides orientation constraints, visual acuity is another related perceptual limit. Normally, human fovea could decompose and process details above its limit visual acuity (1.5-2 degrees of visual angle). It lasts in the frequency range of 0.7-0.5 pixels per degree or 0.33-0.25 cycles per. This range is nearly resembled by the last level low-resolution version of images in usual wavelet decomposition according to sixth column, depth 5, of Table 4.1. Each decomposed level is generated by moving kernels with different window size to any image positions. Spatial frequency of other wavelet analysis levels, varying in accordance with analysing depths, is shown the following the Table 4.1. Spectral energy is usually employed as spectral signature for image collections or individual images. Urban *et al.*[20] analyses different sets of images belonging to four different semantic categories: coast, mountain, street and open-country. Interestingly, the Fourier spectrum of each category has distinguished shape

Depth	0	1	2	3	4	5
Frequency range (cycles per degree)	10.7-5.3	5.3-2.7	2.7-1.3	1.3-0.7	0.7-0.3	0.3-0.2

Table 4.1: Wavelet Levels vs Frequency Range

and frequency range, significantly different from each other [20]. In other words, each general spectral profile and associated distribution histogram of image classes have unique energy distribution. This distribution is proportional to distance d from mean magnitude spectrum normalized by the standard deviation of the category.

$$d(s) = \frac{1}{nm} \sum_{n,m} |s(n,m) - AS(n,m)| \tag{4.3.1}$$

where $AS(n,m)$ is an average spectrum. Carefully observing the spectral profiles could give distinguishing clues for each semantic scene. For example, *"Coastal"* scenes are dominated with horizontal features; therefore, its spectral profiles stretch along vertical axes. Furthermore, spectral profile of *"OpenCountry"* categories is biased toward two upper and lower spectrum. Though almost similar to the *"OpenCountry"* profile, spectrum of *"Street"* images includes more types of features from artificial environment beside horizon-oriented details. Therefore, the diamond of image spectral profile of *"Street"* becomes more significant horizontally. *"Mountain"* categories with its random scenic details have isotropic spectrum while scenes of streets filled with artificial objects have spectrum stretched in both horizontal and vertical axis. From Urban's research, spectral energy distribution seems to be important clues for visually perceptual system of human beings.

Beside image classification, the spectral distribution signature is also useful in visual attention and early visual process. Such energy distribution becomes differentiable clues for features across scales. Spectral profiles of image feature at a particular scale would help differentiate itself from directly upper and lower scale. Let's consider an imaginary experiments with a single square input signal $x(t)$ defined as follows:

$$x(t) = \begin{cases} 1 & t_1 \leq t \leq t_2 \\ 0 & t < t_1 \vee t > t_2 \end{cases} \tag{4.3.2}$$

If $x(t)$ is filtered by a kernel $F()$ with according size (1-D kernel width) $W = \Delta T$ and W is much smaller than non-zero period of the given square signal $\Delta T \lll |t2 - t1|$, the

response will be just two impulse function at $t1, t2$.

$$F(x(t)) = \begin{cases} 1 & t = t_1 \wedge t = t_2 \\ 0 & t \neq t_1 \vee t \neq t_2 \end{cases} \tag{4.3.3}$$

For 2-D signal or image context, the above operation corresponds to a classic edge detection phenomena. Though edges and structures plays important roles in visual perception, their information does not sufficiently represent the whole natural scenes. Natural images are rich with other features like texture, flat regions, etc. besides edges and corners. As mentioned before, image features can be interpreted in terms of energy distribution. For example, edges are the regions of high energy concentration, homogeneous flat regions do not contain much energy while textures, hybrid of edges and flat regions, contains certain amount of energy . If only one window size is used in the analysis, significant responses only come from features or objects which happen to fit into that window size; meanwhile, other meaningful but unfitted features could not be extracted. Therefore, a multi-scale's approach is necessary in order to identify suitable sizes of kernels or fuse features from different scales together. When the mother wavelet is chosen as filtering kernels, window size becomes inversely proportional to frequency range in the Table 4.1. Since choosing a suitable window size is important computation task for spatial feature extraction, WSS proposes a strategy about how spectral density distribution can characterize features at each scale and how the frequency range of processing can be appropriately chosen for multi-scale feature representation. With computation of multi-scale features and appropriate scale selection, WSS is capable of highlighting salient features across scales by using spectral energy distribution.

4.3.2 Energy Distribution in Scale Space and Spectral Space

PSS estimates information from pixel values, time-domain descriptors, by constructing normalized histogram of pixel values as probability distribution.

$$p_h(d) = \frac{n_d}{N} \tag{4.3.4}$$

where p_h is probability of descriptors, the ratio between number of pixels with d descriptors and total image pixels N. Lets use square of pixel d^2 as weights

$$p_e(d) = \frac{n_d * d^2}{\sum_D (n_d * d^2)} = \frac{\rho_x}{E_x} = \frac{\int_{-\infty}^{\infty} (x(t)\delta(x(t) - d))^2 \, dt}{\int_{-\infty}^{\infty} x(t)^2 dt} \tag{4.3.5}$$

Normalized weighted-histogram $p_e(d)$, of signal $x(t) \in L^2(R)$ can obviously be interpreted as ρ_x energy density of descriptors d in time domain. By the isometric property of the Fourier transform, the Probability Distribution Function (PDF) of energy density can be expressed in frequency domain as well.

$$p_e(\hat{d}) = \frac{\rho_{\hat{x}}}{E_{\hat{x}}} = \frac{\int_{-\infty}^{\infty} \hat{x}(f)^2 \delta(\hat{x}(f) - \hat{d}) df}{\int_{-\infty}^{\infty} \hat{x}(f)^2 df} \tag{4.3.6}$$

Or in joint time and frequency domain.

$$p_e = \frac{\rho_{\bar{x}}}{E_{\rho_{\bar{x}}}} = \frac{\rho_{\bar{x}}(t, f)}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \rho_{\bar{x}}(t, f) dt df} \tag{4.3.7}$$

where

$$\rho_{\bar{x}} = \left| \int_{-\infty}^{\infty} x(\tau) g^*_{t,f}(\tau) d\tau \right|^2 \tag{4.3.8}$$

where $\rho_{\bar{x}}$ is energy density in joint time-frequency representation. Pure time descriptors have perfect localization in time, but no localization in frequency, and vice versa for frequency-descriptors. Both extreme time and frequency descriptors make interpretation of constructed PDF, and estimated information difficult to explain. Therefore, it is necessary to find a representation of $g^*_{t,f}(\tau)$, describing spectral density of local energy. For example, Short-Time Fourier Transform (STFT) is the first-known transform capable of generating spectrogram, a graphical representation of local signal energy in time-frequency plan.

$$\rho_x(t, f) = \left| \int_{-\infty}^{\infty} x(\tau) h^*(\tau - t) e^{-2j\pi f \tau} d\tau \right|^2 \tag{4.3.9}$$

STFT identifies spectral density as well as local energy density or information in a short-time period of the signals. However, it does not much benefit scale saliency unless scale parameters are actually considered as in signal description on phase-space. Fortunately, in recent years, alternative scale-based representation, called wavelet-transform (WT) , has been widely addressed among signal processing community, and its fundamental idea is replacing the frequency shifting operation $e^{-2j\pi f\tau}$ by a time (or frequency) scaling operation $\psi(\frac{t-\tau}{a})$, a basic wavelet kernel. Consequently, the energy density in WT framework is formulated as follows.

$$\rho_x(t,a) = \left| \int_{-\infty}^{\infty} x(\tau)\psi(\frac{\tau - t}{a})d\tau \right|^2 \tag{4.3.10}$$

WT coefficients, $\rho_x(t,f)$, averagely measure spectral density of frequency sub-bands, a short range of frequency, in a short period of time. Characteristics of the time-frequency window are specified by two main parameters, time-shift τ and scale a. As derived from short-time spectral representation, a spectrogram, by utilizing scale operations, its distribution of energy density is called scalogram. Given time-scale space, the total signal energy can be rewritten as follows.

$$E_{\rho_x} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \rho_x(t,a)dtda \tag{4.3.11}$$

and probability of time-scale descriptors can be specified with scale parameters

$$P_e(t,a) = \frac{\rho(t,a)}{E_{\rho_x}} \tag{4.3.12}$$

Generally, the wavelet transform can help generate frequency sub-band coefficients, square of which over total energy are density distribution of that sub-band in time-scale space.

4.3.3 Wavelet Transform

The discussion in the Section 4.3.1 about usefulness of time-frequency-scale representation and its corresponding time-frequency energy distribution suggests that wavelet-representation may be ideal candidates for the investigation into distribution of energy density and other statistical property across multiple scales of natural images. During a quite short history of wavelet analysis, this research field has been very fruitful with several analysing techniques and wide range of characteristics. In this chapter ,

only standard techniques such as discrete wavelet transform (DWT), discrete wavelet packet transform (DWPT), quaternion wavelet transform with best-basis(QWTBB), and quaternion wavelet packet transform with best-basis (QWPTBB). are deployed as potential descriptors. This Sub-Section first summarizes discrete and real wavelet and then wavelet packet transform with best basis (DWT,DWPTPP) for its theoretical background; then QWT and QWPTBB, quaternion versions of two prior wavelet transforms.

Discrete Wavelet Transform (DWT)

Though wavelets were firstly introduced in the early 20th century by Alfred Harr, they are only developed rapidly much later. Only until recently, they have been widely employed in many computer vision problems such as image or video de-noising, enhancement, coding, and pattern classification [126–128] . Signal analysis for frequency components can be achieved by Fourier transform (FT) but FT does not provide suitable tool for time-frequency analysis of images. Short-time Fourier Transform (STFT) is an extension from FT approach for analysing local frequency analysis at a short period of time [129]. Noteworthy that, STFT can be used for taking the spatial interval in 2-D signal instead of time period in 1-D type since there is no time dimension for still images. However, STFT utilizes fixed window kernels for every data blocks across input signals; this property makes STFT less suitable for complex signal analysis, especially signals with strong semantic structures appearing across multiple scales. In other words, STFT only succeeds with signals whose features are embedded in fixed definite temporal or spatial regions or there is prior knowledge about a suitable size of window kernel for STFT processing. Without the above conditions, STFT would totally miss signal features. Theoretically, the disadvantage can be avoided by employment of STFT with multiple kernel sizes; however, it raises up other issues such as what range of sizes would be chosen to optimally extract useful features with reasonably computational effort.

Problems of STFT in analysing local frequency have motivated development of multi-scale wavelet techniques for better local frequency representation. Since limitations of STFT is due to fixed-size processing windows, wavelet analysis deploys multi-resolution filter-banks on input signals. As illustrated in Figure 4.2, 1-D signals are decomposed into low-pass and high-pass components. In case of 2-D input signals, the filtered outputs are four sub-bands: low-low, high-low, low-high, and high-high in regards of processing orientation. Intuitively, 2-D signals analysis includes row-wise 1-D analysis followed by column-wise 1-D analysis or vice verse. With respect to processing



(a) Sample Image **(b)** Wavelet Decomposition

Figure 4.1: Wavelet analysis of a sample image

direction, high-low sub-bands tend to extract horizontal features, low-high sub-bands prefer vertical features, high-high sub-band detects diagonal features, and low-low are low-resolution approximation of original signals by inverse dyadic scale. Let's assume that input signals are two-dimensional grey-scale image $f(x,y)$, and the scaled mother wavelets have followed mathematical form$\psi_{s,i}(x,y) = 2^s\psi(2^s x, 2^s y)|_{i=\{v,h,d\}}$ for vertical, horizontal and diagonal sub-bands and a scaling function $\phi_S(x,y) = 2^S\phi(2^S x, 2^S y)$ for low-resolution signals with $s \geq S$. Then, any images $f(x,y) \in L_2(\mathbb{R})$ can be represented as.

$$f(x,y) = \sum_{x,y} c_S(x,y)\phi_S(x,y) + \sum_{i=\{v,h,d\}} \sum_{x,y,s\geq S} d_{s,i}(x,y)\psi_{s,i}(x,y) \tag{4.3.13}$$

where (4.3.14)

$$c_S(x,y) = \int f(x,y)\phi_S(x,y)dxdy \tag{4.3.15}$$

$$d_{s,i}(x,y)|_{i=\{v,h,d\}} = \int f(x,y)\psi_{s,i}(x,y)dxdy \tag{4.3.16}$$

$c_S(x,y)$ and $d_{s,i}(x,y)$ are scaling coefficients and wavelet coefficients from vertical, horizontal and diagonal sub-bands. The parameter S represents the lowest analysing depth while s is higher decomposing levels in multiple scale-space framework. As mentioned before, 2-D DWT can be obtained by tensors products of 1-D DWT when the orginal image $f(x,y)$ is analysed along two dimensions x and y separately. As a result, the scale function $\phi(x,y)$ is approximated as $\phi(x)\phi(y)$ and filter-banks of three directional sub-bands are $\phi(x)\psi(y), \psi(x)\phi(y), \psi(x), \psi(y)$. In the Figure 4.2, discrete wavelet transforms are carried on the sample image in the left hand-side. On the right-hand

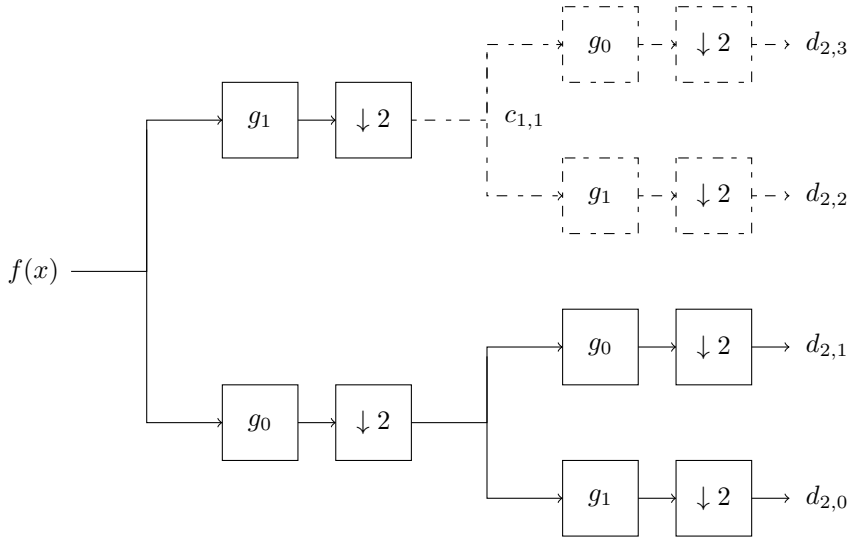


Figure 4.2: Wavelet (solid line) & Wavelet Packet Decomposition (solid and dash lines) for 1-D signal

side contains decomposed results by two levels with three distinctive sub-bands and a down-sampled version of the original image. Real-value wavelet transform like DWT suffers from shift-variance, a small shift in the signal can greatly change magnitudes of wavelet coefficients around singularities. Furthermore, it has no phase to embed signal location information; therefore, aliasing effects would be introduced into recovery process. These issues need serious consideration whenever the discrete real wavelet transform is employed. Therefore, modelling statistical property of DWT coefficients' magnitude across scales might request extra investigation with those draw-backs in mind. Further arguments and details about this matter of DWT descriptors will be discussed in the Section 4.2.

Discrete Wavelet Packet Transform Best Basis (DWPTBB)

Well-known DWT can be computed efficiently by orthonormal finite-infinite-response (FIR) conjugate quadrature filter banks g_0, g_1 including analysis low-pass and high-pass filters, denoted as ϕ_S and $\psi_{s,i}$ respectively. The low-pass coefficients $c_S(x, y)$ are decomposed recursively for a number of levels, and inverse discrete wavelet transform is calculated by an inverse filter bank. The extension from normal wavelet transform (DWT) to wavelet packets transform (DWPT) is straightforward by an additional step at each processing level. Instead of decomposing only low-pass coefficients in the Low-Low sub-band for 2D input signals, the transform performs decomposition of high-pass coefficients in Low-High, High-Low and High-High sub-bands as well. As a result, all coefficients of DWPT can be neatly arranged in a binary tree and addressed

as follows.

$$d_{s,i}(x,y)\ , s \in [0,S], i \in [0,4^s-1] \tag{4.3.17}$$

where s is an analysis depth in the tree, S denotes the deepest decomposed level, and i is the node index in this depth. With regards to other representations, wavelet packets have advantages in their adaptability to varying statistical structure. Unlike Fourier Transform with one fixed-size base or normal wavelet transform with a fixed number of bases, the "best" orthonormal bases from dictionary of basis can be acquired after wavelet package decomposition in a best basis approach. This idea is initially proposed by Coifman *et al.*[130] mainly for signal compression. Therefore, this "best" basis is the best in terms of compressing ratio which often desires the sparsest representation. In other words, input signals can be characterized by fewest large coefficients. Supposed the whole best basis operation is denoted as B^2 which exhaustively goes through the whole binary tree to look for locations of a set basis with parameters $(\mathbf{s},\mathbf{i})$ such that the there is a minimum amount of uncertainty measured by Shannon entropy. More details can be found in Coifman's works [130], and B^2 can be summarized mathematically as follows.

$$B^2(d_{s,i}) : (\mathbf{s},\mathbf{i}) = argmin_{s,i}(\sum H(d_{s,i})) \tag{4.3.18}$$

Sometimes, "brute-force-attack" every branch of the tree is not possible or preferable due to intensive requirement of computational power. Fortunately, there exist fast algorithms to implement the best basis for given signals [130]. Then, the optimum time-frequency representation can be achieved by tilting the time-frequency plan in accordance with best-basis algorithms. Though the representation may be optimally sparsest in time-frequency domain, whether sparseness of features suitably matches performance of human visual attention is still a question to be answered. Experiments have been carried out and performance comparison between the DWT case and DW-PTBB case is reported in the Section 4.4.

Quaternion Wavelet Transform (QWT)

Like wavelet packet transform in the previous discussion, Quaternion Wavelet Transform (QWT) is extended a nd enhanced to eliminate shift-variance problems from real discrete wavelet transform (DWT). Though there are a few different definitions and implementations of Quaternion Wavelet [128, 131], the QWT implementation in this paper is inspired by Chan's research *et al.*[128]. Some background about complex

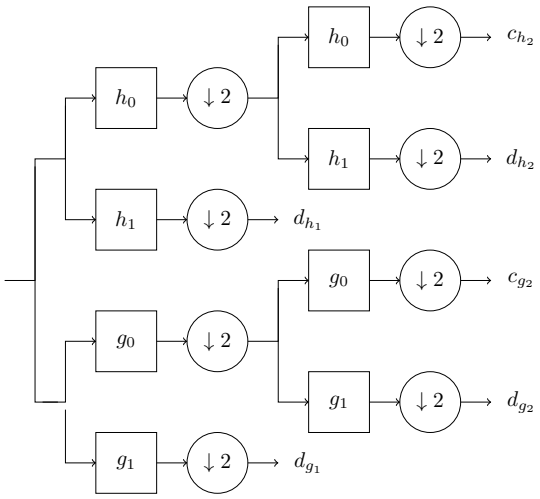


Figure 4.3: Illustration of DTCWT

wavelet transform (CWT) with its implementation dual-tree complex wavelet transform (DTCWT) Kingsbury *et al.*[132] need to be reviewed before the QWT can be explained and discussed. In discrete wavelet transform, 2-D DWT can be considered as concatenation of two consecutive 1-D DWTs. Though the same process does not exactly happen in 2-D DTCWT or QWT in a straightforward manner, a similar concept is used to easily explain how QWT can be computed. It means 1-D DTCWT will be elaborated first; then, 2-D QWT signals are discussed as well as how it may handle processing along different orientations and sub-bands.

Real-value DWT has well-known drawbacks in terms of shift-invariance and phases to encode coefficient locations. Kingsbury *et al.*[133] reckons these problems and proposes a dual-tree CWT as a specific solution. The rationale behind the dual-tree approach is the usage of complex numbers for wavelet coefficients for tackling the DWT problem of shift-variance. Furthermore, complex extension of wavelet transform makes phase extraction from wavelet coefficients possible since complex wavelet transforms have both real and imaginary values unlike real DWT with only one real value for each coefficient. Real and imaginary components of the dual-tree CWT are generated by two sets of wavelet and scaling functions $\psi_h, \psi_g$and ϕ_h, ϕ_g. Moreover, filter-banks $h_0, h_1$and g_0, g_1 have to be independent and orthogonal as shown in Figure 4.3.

The notations $\phi_h(x)$ and $\psi_h(x)$ are denoted for scaling and wavelet functions corresponding to filter-banks h_0, h_1. In addition, c_{h_S} and $d_{h_s}$with $s \leq S$ denotes first set of DTCWT coefficients. Similar notations are used for the second set of scaling and wavelet functions $\phi_g(x)$ and $\psi_g(x)$ with filter banks g_0, g_1 and according coefficients c_{g_S} and d_{g_s} with $s \leq S$. Wavelet functions $\psi_h(x)$ and $\psi_g(x)$ form two binary trees in Figure 4.3; moreover, at leaves of each tree are the real and imaginary parts of complex

analytic wavelet coefficients.

$$\psi^{(c)}(x) = \psi_h(x) + j\psi_g(x) \tag{4.3.19}$$

Moreover, the imaginary wavelet $\psi_g(x)$ is 1-D Hilbert Transform of the real-wavelet $\psi_h(x)$:

$$\psi_g(x) = \mathcal{HT}(\psi_h(x)) \tag{4.3.20}$$

Any complex wavelet coefficients are formed by two correspondent real-value wavelet transforms; therefore, a complex-value wavelet generates redundant tight frames of which numbers are twice as many as a real-value wavelet does. This redundancy in complex wavelet frames prevents non-oscillating magnitudes of coefficients around singularity points as well makes the transform near-shift invariant. Furthermore, there is no energy (or little energy in practice) in the negative region of frequency because of a relationship between two wavelet functions ψ_h and ψ_g.

$$\psi_g(\omega) = \begin{cases} -j\psi_h(\omega) & \omega > 0 \\ j\psi_h(\omega) & \omega < 0 \end{cases} \tag{4.3.21}$$

then

$$\begin{aligned} \psi^c(\omega) =& \mathcal{F}(\psi^c(x)) && (4.3.22)\\ =& \mathcal{F}(\psi_h(x) + j\psi_g(x)) && (4.3.23)\\ =& \psi_h(\omega) + j\psi_g(\omega) && (4.3.24) \end{aligned}$$

and

$$\psi^c(\omega) = \begin{cases} 2\psi_h(\omega) & \omega > 0 \\ 0 & \omega < 0 \end{cases} \tag{4.3.25}$$

Thus, the Fourier transform of complex wavelet coefficients $\psi^c(\omega)$ has no energy in the negative frequency region, which makes DTCWT an analytic wavelet transform with analytic output signals. Due to this analyticity, the dual-tree wavelet transform has implicitly managed to include all information in the half positive plan of the frequency domain.

It is quite straight forward for 2-D DWT expansion from 1-D DWT, discussion in the previous Sub-Section 4.3.3. However, it is unfortunately not easy for similar expan-

sion from 1-D DTCWT to 2-D DTCWT transforms because Hilbert Transform (HT) and analytic signals need a theoretical extension for 2-D signals. Furthermore, there exists not only one but several definitions for different zero-out regions (negative frequency domain in 1D DTCWT) , signal-power regions (positive frequency domain in 2D DTCWT). However, only Bulow definition [131] of analytic quaternion signals which combines both partial and total Hilbert transform (HT) is considered here. Partial HTs are done along either x or y directions only; meanwhile, total HT is carried out on both directions simultaneously. They are defined as :

$$f_{\mathcal{H}_{i_1}}(\mathbf{x}) = f(\mathbf{x}) \circ \frac{\delta(y)}{\pi x} \tag{4.3.26}$$

$$f_{\mathcal{H}_{i_2}}(\mathbf{x}) = f(\mathbf{x}) \circ \frac{\delta(y)}{\pi x} \tag{4.3.27}$$

$$f_{\mathcal{H}_i}(\mathbf{x}) = f(\mathbf{x}) \circ \frac{1}{\pi^2 xy} \tag{4.3.28}$$

The $f_{\mathcal{H}_{i_1}}, f_{\mathcal{H}_{i_2}}(\mathbf{x})$ are partially Hilbert transformed along x and y axis consequently, and $f_{\mathcal{H}_i}(\mathbf{x})$ is total HT; while $\circ$ denotes 2-D convolution. Each 2-D CWT basis is a complex analytic function, computationally equivalent to a product of two 1-D complex wavelet functions either along only one or both axis. Similar to expansion of discrete real wavelet, the diagonal wavelet sub-band is defined as $f(\mathbf{x}) = \psi_h(x)\psi_h(y)$. Other total and partial HT are products of coefficients from different sets of wavelet functions deployed in the 1-D CWT implementation.

$$(f_{\mathcal{H}_{i_1}}, f_{\mathcal{H}_{i2}}, f_{\mathcal{H}_i}) = (\psi_g(x)\psi_h(y), \psi_h(x)\psi_g(y), \psi_g(x)\psi_g(y)) \tag{4.3.29}$$

To unify all different Hilbert Transform in a meaningful and compact representation, quaternion algebra and treat $f(\mathbf{x})$ as a real part and $(f_{\mathcal{H}_{i_1}}, f_{\mathcal{H}_{i2}}, f_{\mathcal{H}_i})$ can be utilised as three imaginary components [128].

$$f_A^q(\mathbf{x}) = f(\mathbf{x}) + j_1 f_{\mathcal{H}_{i_1}}(\mathbf{x}) + j_2 f_{\mathcal{H}_{i_2}}(\mathbf{x}) + j_3 f_{\mathcal{H}}(\mathbf{x}) \tag{4.3.30}$$

More details about theory behind QWT and its special characteristics such as its singular cases, three phases, and zero-out regions can be found in Chan etal 's and Bulow 's publications [128, 131]. According to the above transformation of quaternion wavelet, four quadrant components of 2-D wavelet $(f\ , f_{\mathcal{H}_{i_1}}, f_{\mathcal{H}_{i2}}, f_{\mathcal{H}_i})$ can be organized into a quaternion number; for example, following is a diagonal signal with quadrant components.

$$(f\ , f_{\mathcal{H}_{i_1}}, f_{\mathcal{H}_{i2}}, f_{\mathcal{H}_i}) = (\psi_h(x)\psi_h(y), \psi_g(x)\psi_h(y), \psi_h(x)\psi_g(y), \psi_g(y)\psi_g(y)) \tag{4.3.31}$$

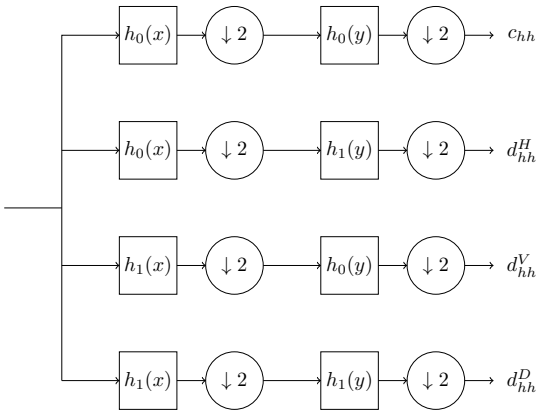


Figure 4.4: Illustration of dual-tree complex wavelet transform

There is a diagonal quaternion wavelet functions for the correspondent sub-band, which is mathematically defined as follows.

$$\psi^D(x,y) = \psi_h(x)\psi_h(y) + j_1\psi_g(x)\psi_h(y) + j_2\psi_h(x)\psi_g(y) + j_3\psi_g(x)\psi_g(y) \tag{4.3.32}$$

To compute the QWT coefficients, Kingsbury proposed a separable 2-D implementation [133] of dual-tree filter-banks previously illustrated in Figure 4.3. At each filtering stage, both two-sets of wavelet filters h and g are independently applied to each dimension x and y of a 2-D image. For example, the filter-bank h is applied along both axis; then, it yields the scaling coefficients c_{hh_s} and three diagonal,vertical and horizontal wavelet coefficients $d^D_{hh_s}, d^V_{hh_s}$ and $d^H_{hh_s}$ respectively as shown in the Figure 4.4. Dual-tree implementation of two separated filter-banks for 1-D signal can be considered as four independent filter banks for 2-D signals according to all possible combinations of filter for one dimension (hh, hg, gh, gg). With these combinations of filters and corresponding wavelet functions $\psi(x)\phi(y), \phi(x)\psi(y)$ and $\psi(x)\psi(y)$ are generated four components of quaternion wavelet transform for horizontal, vertical, and diagonal sub-bands. Four different wavelet coefficients from these filter banks are arranged by quaternion algebra to obtain QWT coefficients. For example, a coefficient from diagonal wavelet sub-band of QWT can be written in terms of responses from independent filter-banks as follows.

$$d^D_s = d^D_{hh} + j_1 d^D_{gh} + j_2 d^D_{hg} + j_3 d^D_{gg} \tag{4.3.33}$$

So far, a diagonal sub-band has been taken as example for showing how QWT can be computed. The construction and properties for other two sub-bands are similar to what have been done for diagonal sub-bands. Except that the axis combinations results in a horizontal sub-band $\psi(x)\phi(y)$ or a vertical sub-band $\psi(x)\psi(y)$ instead of a diago-

nal sub-band $\psi(x)\psi(y)$. In summary, QWT at each stage sports three quaternion sets corresponding to three sub-bands; each quaternion contains four wavelet functions. Therefore, there are 12 functions in total which can be easily seen as matrix of functions as follows.

$$\begin{bmatrix} d_s^H & d_s^V & d_s^D \end{bmatrix} \overset{q}{=} \begin{bmatrix} d_{hh}^H & d_{hh}^V & d_{hh}^D \\ d_{gh}^H & d_{gh}^V & d_{gh}^D \\ d_{hg}^H & d_{hg}^V & d_{hg}^D \\ d_{gg}^H & d_{gg}^V & d_{gg}^D \end{bmatrix}$$
$$= \begin{bmatrix} \psi_h(x)\phi_h(y) & \phi_h(x)\psi_h(y) & \psi_h(x)\psi_h(y) \\ \psi_g(x)\phi_h(y) & \phi_g(x)\psi_h(y) & \psi_g(x)\psi_h(y) \\ \psi_h(x)\phi_g(y) & \phi_h(x)\psi_g(y) & \psi_h(x)\psi_g(y) \\ \psi_g(y)\phi_g(y) & \phi_g(y)\psi_g(y) & \psi_g(y)\psi_g(y) \end{bmatrix}$$

Columns of the above matrix correspond to the quaternion wavelet functions of the horizontal sub-band d^H, the vertical sub-band d^V, and diagonal sub-band d^D from left-to-right respectively. The three according wavelet coefficients are d_s^H, d_s^V and d_s^D and the $\overset{q}{=}$ operators mean formation of quaternion number by coefficients along each column. Though coefficients of quaternion wavelets possess rich phase information, this research currently focuses on magnitudes of each wavelet sub-band. Therefore, magnitudes of horizontal, vertical sub-bands can be computed according to formula of quaternion magnitudes as follows.

$$\|d_s^H\| = \sqrt{(d_{hh}^H).^2 + (d_{gh}^H)^2 + (d_{hg}^H).^2 + (d_{gg}^H)^2}$$
$$\|d_s^V\| = \sqrt{(d_{hh}^V).^2 + (d_{gh}^V)^2 + (d_{hg}^V).^2 + (d_{gg}^V)^2}$$
$$\|d_s^D\| = \sqrt{(d_{hh}^D).^2 + (d_{gh}^D)^2 + (d_{hg}^D).^2 + (d_{gg}^D)^2}$$

While the final approximated version of input signals, which are not decomposed further by the transform, have its magnitude computed by quaternion algebra.

$$\|c_S\| = \sqrt{(c_{hh}).^2 + (c_{gh})^2 + (c_{hg}).^2 + (c_{gg})^2}$$
$$= \sqrt{(\phi_h(x)\phi_h(y)).^2 + (\phi_g(x)\phi_h(y))^2 + (\phi_h(x)\phi_g(y)).^2 + (\phi_g(y)\phi_g(y))^2}$$

Quaternion Wavelet Packet Transform Best Basis (QWPTBB)

To construct a packet form of QWT (QWPT), each and every sub-band c_S, d^H, d^V, d^D should be repeatedly decomposed by low-pass (h_0, g_0) and high-pass filters (h_1, g_1).

Bayram *et al.*[134] has investigated into formation of wavelet packets for DT-CWT, an equivalent form of QWT. In order to get an analytic wavelet packet of quaternion values, filter banks need to be chosen in a specific way such that the Hilbert transform relationship is preserved. In Bayram's works [134], the analytic wavelet transformed can be achieved if whatever filter-bank is used to decompose the first filter-bank of QWT should also be used for the second (dual) filter-banks. Another important point about the extension to QWPT is the choice of the extension filters $f_i(x)$. It has been found that the only necessary constrain to preserve the Hilbert transform property is forcing the usage of the same filter-pairs $f_0(x), f_1(x)$ in both filter-banks of QWT or DT-CWT . Therefore, any Conjugate Quadrature Filter (CQF) pair of filter-banks with short support, frequency selectivity or possessing a number of vanish moments can be candidates for the extension filter. Noted that, the above criteria such as CQF pair of filters have been employed for extending a regular DWT. Like other derivatives of DT-CWT or QWT, the quaternion wavelet packet transform (QWPT) are approximately shift-invariant, which means the energy in each sub-band is approximately preserved if the input signals are shifted by a number of samples. Noteworthy that, there are other methods beside QWPT with shift-invariant property in wavelet packet decomposition. For example, by performing an exhaustive search over all shifted wavelet packet bases to find the "best basis" according to a certain cost function [135], the orthonormal wavelet packet transform becomes shift-invariant in a sense that energy in each sub-band is invariant to transition of input signals . This approximate shift-invariance becomes very useful and important in search for suitable energy descriptors; moreover, it guarantees that DT-CWT, DT-CWPT or QWT, QWPT would have energy descriptors, robust to certain amount of affine transformation in input signals. Interactions of filtering both low and high components at each stage of DT-CWT introduce complete structures of all possible sub-bands, generated by the filter-bank pair. Each tree forms a unique frequency profile of input signals, and among those countless numbers of possibilities, there exists a frequency decomposition being more sparse and compact than the others. It is called the best-basis, representing the input signals with fewest wavelet coefficients. A fast algorithm for identifying such best basis has been reported in extension from DWT to DWPT [130]. The strategy, mentioned in the Sub-Section 4.3.3, can be adopted for searching best-basis in QWT. In brief, it is looking for a path in a binary of decomposition to minimize a Shannon entropy cost function; more details can be found in the work of Coifman *et al.*[130]. After "best basis" searching for QWPT decomposition, magnitudes and energy of coefficients can be identified at a specific location by a simple quaternion algebra. $\|q(a,b,c,d)\| = \sqrt{a^2+b^2+c^2+d^2}$.

4.3.4 Wavelet Coefficients Correlation

Section 4.3.2 has discussed the potential of using distribution of energy density, localized time-scale element or wavelet elements instead of pixel-value probability distribution. Only general 1-D signal has been considered so far and these elements are assumed to be independent or at least linearly uncorrelated; however, this assumption only works for random variables as input signals.

Practically, except total noise, any meaningful signals often has specific structures persistent across multiple time-scale element in 1-D case. For 2-D signals like natural images, an additional orientation needs considering; in other words, their wavelet coeffcients are highly statistically related across scales, orientation, spaces. This phenomenon is systematically studied and confirmed in Azimifar *et al.*research [6]. The authors have conducted an empirical study of joint wavelet statistics for texture and natural images to investigate correlation relationship between neighbouring coefficients. Examination of these dependencies helps propose appropriate models for such a transform-domain algorithm. Though Azimifar's work [6] covers most of linear dependencies and a bit of non-linear relations, its proposals are evaluated on a collection of 5000 real images. Therefore, persistence of structures across multiple scales is generally true at least for natural images, the main researching objects. In brief, there exists a few elementary correlation relationships as follows.

- The spatially-localized and sparse correlation structure have a clear persistence across scales.
- Every coefficient exhibits correlations extending across multiple scales, with spatially near neighbours both within and across orientations.
- A subband coefficients at the same spatial locations but from different orientations are not linearly correlated.
- Within-subband, inter-subband, and inter-scale correlations are highly oriented and persistent across local neighbors of its parent.

Figure 4.5 clearly illustrates all correlations of wavelet coefficients, their preferences to locality must be expected. It increases toward finer scales, which supports persistency property of wavelet coefficients. A single coefficient correlates with its parents as well as neighbours across orientations and scales. Among these dependencies, the most vital findings for this work are uncorrelated sibling coefficients across orientation and strong correlated coefficients across scales since it theoretically allows uncertainty

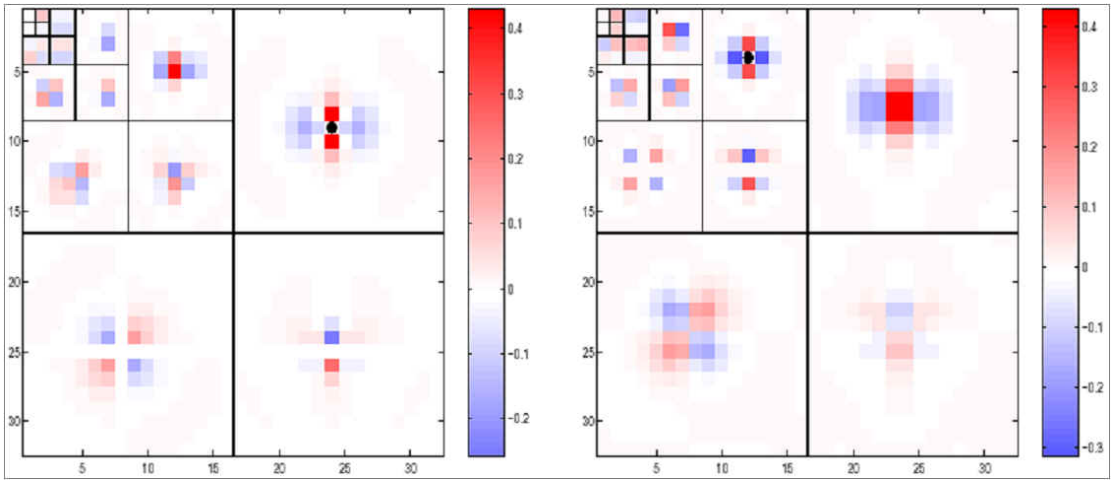


Figure 4.5: Inter-band and intra-band correlation of wavelet coefficients [6]

and mutual information estimation of a 2D time-scale element or a wavelet sub-band energy descriptor. To elaborate this point, lets consider two adjacent scales $s1, s2$ and their corresponding coefficients $w_i(f(x, y, s))$ of horizontal, vertical, diagonal orientation $i = v, h, d$ for 2-D signals $f(x, y)$. Due to non-correlation of sibling coefficients across orientation, it is possible to consider three wavelet coefficients as a multivariate variable $\mathbf{w}_s = (w_h, w_d, w_v)$ with uncertainty estimation by distribution of energy density across three orientations $H(\mathbf{w}_s)$. For two adjacent sub-band s_1, s_2, there are two multivariate variables $\mathbf{w}_{s_1}$ and $\mathbf{w}_{s_2}$ corresponding two entropy values $H(W_{s_1})$ and $H(\mathbf{w}_{s_2})$, and the mutual information between two variables due to inter-scale dependencies between correspondent wavelet coefficients are computed as follows.

$$MI(\mathbf{w}_{s_1}, \mathbf{w}_{s_2}) = H(\mathbf{w}_{s_1}, \mathbf{w}_{s_2}) - H(\mathbf{w}_{s_1}) - H(\mathbf{w}_{s_2}) \tag{4.3.34}$$

From that basic observation about inter-scale and intra-scale wavelet coefficients of natural images, the core idea is developed. More details about sub-band energy descriptors and how to measure their uncertainty and mutual information will be clearly explained in the Section 4.3.4.

Wavelet Sub-band Energy Descriptor

Interesting relationship between basis-project methods and scale saliency are repeatedly discussed in several publications [22], [53]. Kadir [22] actually discusses about behaviours of non-saliency and saliency regions in spectral and wavelet domain. A simple, flat, non-salient regions or images are sufficiently described by a single wavelet sub-band; meanwhile, complicated data and structure regions require more wavelet sub-bands. This flexibility makes wavelet sub-bands as potential alternative descrip-

tors for scale saliency. Like pixel-value descriptors, real wavelet sub-bands must be treated as discrete variables due to its theoretical restriction, data-analysis uncertainties, $\sigma_t\sigma_\omega \geq \frac{1}{2}$. In other words, it is impossible for continuous wavelet sub-bands distribution at any specific location. With respect to available mathematical definition of PSS [136] for discrete pixel descriptors, a mathematical model of WSS is defined with discrete wavelet sub-band descriptors, $\{e \in \mathbf{E}, \mathbf{E} = \{e_1, e_2, \ldots e_i\}\}$, in Equations 4.3.35, 4.3.36, 4.3.37, 4.3.38 whereof e, E are an element and a set of sub-band descriptors consecutively.

$$Y_E(s_p, \vec{x}) \triangleq H_E(s_p, \vec{x}) W_E(s_p, \vec{x}) \tag{4.3.35}$$

$$H_E(s, \vec{x}) \triangleq -\sum_{e \in E} p_E(e, s, \vec{x}) log_2 \left(p_E(e, s, \vec{x})\right) \tag{4.3.36}$$

$$W_E(s, \vec{x}) \triangleq \frac{s^2}{2s-1} \sum_{e \in E} |p_E(e, s, \vec{x}) - p_E(e, s-1, \vec{x})| \tag{4.3.37}$$

$$\vec{s_p} \triangleq \{s : H_E(s-1, \vec{x}) < H_E(s, \vec{x}) > H_E(s+1, \vec{x})\} \tag{4.3.38}$$

Where WSS scale-saliency, Y_E, at a chosen scale s_p at each image location $(\vec{x})$ is defined in Equation 4.3.35 as product between feature-space saliency H_E and inter-scale saliency W_E. Feature-space saliency of WSS, $(H_E(s, \vec{x}))$, is defined as Shannon entropy of a discrete wavelet sub-band descriptor (e) in Equation 4.3.36. Inter-scale saliency W_E is defined as differences between PDF of e descriptors of two consecutive scales s and $s-1$. Finally, the scale s_p is defined in Equation 4.3.38 as a scale where feature-space saliency H_E is maximized. It is noteworthy that Equations 4.3.35-4.3.38 represents only definitions of WSS scale saliency and relevant measurements but not its actual computation. Moreover, definition of DIS, a WSS's derivative only differs from WSS in how s_p is selected. DIS selects the scale s_p such that the inter-scale W_E is maximised across scales. Mathematically, it is written as follows

$$\vec{s_p} \triangleq \{s : W_E(s-1, \vec{x}) < W_E(s, \vec{x}) > W_E(s+1, \vec{x})\} \tag{4.3.39}$$

The computational frameworks of WSS and DIS are only achieved when descriptors e are mathematically defined. Let's consider 2-D DWT, summarized in Sub-Section 4.3.3, with three sub-bands vertical (v), horizontal (h) and diagonal (d) sub-bands at each particular dyadic scale s represented by three sets of wavelet coefficients $(\mathbf{w_i})$ accordingly in Equation 4.3.40. The coefficients $\mathbf{w_i}$ is utilized in Equation 4.3.41 to compute

DWT-based sub-band energy as descriptors (e_{DWT}) for WSS.

$$\mathbf{w_i}(f(x,y,s_j))|_{i=\{v,h,d\},s_j=\{s_1,s_2,\ldots,s_n\}} = f(x,y) * \psi_{s,i}(x,y,s_j) \tag{4.3.40}$$

$$e_{DWT} \triangleq \mathbf{P}\{\mathbf{w_i}(f(x,y,s_j))\}|_{i=\{v,h,d\},s_j=\{s_1,s_2,\ldots,s_n\}} = |\mathbf{w}(f(x,y,s_j))|^2 \tag{4.3.41}$$

Where f represents image, x, y, s_j are locations and scales of wavelet analysis $\psi_{s,i}(x,y,s_j)$ represents the DWT mother wavelet and P is energy of wavelet coefficients. The DWT sub-band descriptor e_{DWT} is usually more sparse than the pixel-based descriptor. For example, decomposition of an image with DWT for 4 or 5 levels generates 12 or 15 sub-bands descriptors while there are always 255 pixel-based descriptors for a grey-scale image.

Besides DWT, other types of wavelet analysis could also be utilized as well; for example, discrete wavelet packet with best basis analysis (DWPTBB), summarized in Sub-Section 4.3.3. The full wavelet packet transform breaks signals into sub-bands with the same bandwidth and the same dyadic scale. It would not fit into the scale-saliency concept, requiring descriptors at different scales. Fortunately, this full wavelet packet tree usually over-describes image properties, and such tree can be optimized by Best Basis (B^2) finding operation. The optimized wavelet packet tree often has projected basis across dyadic scales since some small image details are best described with a basis at finer resolution while other big details prefer another basis with coarser resolution. The DWPTBB coefficients are utilized for computation of the DWPTBB-based sub-band descriptor, e_{DWPTBB}, in Equations 4.3.42, 4.3.43.

$$\mathbf{w_i}(f(x,y,s_j))|_{(i,s_j)=B^2(\mathbf{w_i}f(x,y,s_j))} = f(x,y) * \psi_{s,i}(x,y,s_j) \tag{4.3.42}$$

$$e_{DWPTBB} \triangleq \mathbf{P}\{\mathbf{w_i}(f(x,y,s_j))\}|_{(i,s_j)=B^2(\mathbf{w_i}f(x,y,s_j))} = |\mathbf{w_i}(f(x,y,s_j))|^2 \tag{4.3.43}$$

Where f represents image, x, y, s_j are locations and scales of wavelet analysis, $B^2(\mathbf{w_i}f(x,y,s_j))$ represents best-basis of this wavelet analysis, $\psi_{s,i}(x,y,s_j)$ represents the DWPTBB mother wavelet, and P is energy of wavelet coefficients. Comparison between mathematical statements 4.3.40,4.3.41 and 4.3.42,4.3.43 for sub-bands descriptors of Discrete Wavelet Transform (DWT) and Discrete Wavelet Packet Transform Best Basis (DWPTBB) highlights fundamental differences between these descriptors. While DWT provides determinant basis-projection methods with pre-computed basis and fixed structure of sub-bands, DWPTBB is driven by image data. Then, its number and structure of sub-bands are specified by Best Basis (B^2) finding operator [130]. It requires more operations; however, more faithful and adaptive descriptors can be achieved.

The biggest disadvantage of DWTBB and DWT is depending on shift-variant real dis-

crete wavelet transforms. It means that projection of coefficients not only depends on data but also its relative location on scenes. As shown in Equation 4.3.44, wavelet coefficients are differed with respect to their locations (x, y), $(x + \delta(x), y + \delta(y)$ though the image signal f does not change. Hence, the wavelet-based descriptors $e_{DWT}, eDWPTBB$ are also shift-variant.

$$\mathbf{w_i}(f(x, y, s_j)) \neq \mathbf{w_i}(f(x + \Delta(x), y + \Delta(y)), s_j) \quad , \exists x, y, s_j, \mathbf{w_i}, \Delta(x), \Delta(y) \tag{4.3.44}$$

$$\mathbf{P}\{\mathbf{w_i}(f(x, y, s_j))\} \neq \mathbf{P}\{\mathbf{w_i}(f(x + \Delta(x), y + \Delta(y), s_j))\}, \exists x, y, s_j, \mathbf{w_i}, \Delta(x), \Delta(y) \tag{4.3.45}$$

As the fourth criterion for good information measurement of Starck *et al.*[120] states that entropy must work in the same way regardless of descriptors' locations; in other words, the descriptors should be linear shift-invariant. Both DWT-based and DWTBB-based sub-band descriptors do not satisfy that condition. Utilisation of these shift-variant descriptors might lead to different information estimation for identical data at two different locations. The shift-variance of real-wavelet transform can be avoided by design of complex wavelet transform; for instances, quaternion wavelet transform (QWT) [137] [128], summarized in Sub-Section 4.3.3 and quaternion wavelet packet transform with best-basis (QWPTBB) [134], summarized in Sub-Section 4.3.3. In other words, QWT and QWPTBB are shift-invariant derivative of DWT and DWPTBB. General formula of complex coefficients and their corresponding wavelet-based sub-band descriptors, e_{QWT} and e_{QWPTBB}, are described in Equations 4.3.46,4.3.47.

$$\mathbf{w_i}(f(x, y, s_j))|_{i=\{v,h,d\}, s_j=\{s_1,\ldots s_n\}} = f(x, y) * \left(\psi_{g,s,i}(x, y, s_j) + j\psi_{h,s,i}(x, y, s_j)\right) \tag{4.3.46}$$

$$e_{QWT} \triangleq \mathbf{P}\{\mathbf{w_i} f(x, y, s_j)\}|_{i=\{v,h,d\}, s_j=\{s_1,\ldots,s_n\}} = \|\mathbf{w_i}(f(x, y, s_j))\|_2^2 \tag{4.3.47}$$

$$\mathbf{w_i}(f(x, y, s_j))|_{(i,s_j)=\{B^2(\mathbf{w_i} f)\}} = f(x, y) * \left(\psi_{g,s,i}(x, y, s_j) + j\psi_{h,s,i}(x, y, s_j)\right) \tag{4.3.48}$$

$$e_{QWPTBB} \triangleq \mathbf{P}\{\mathbf{w_i} f(x, y, s_j)\}|_{(i,s_j)=B^2(\mathbf{w_i} f)} = \|\mathbf{w_i}(f(x, y, s_j))\|_2^2 \tag{4.3.49}$$

Where f represents image, x, y, s_j are locations and scales of the wavelet analysis, $B^2(\mathbf{w_i})$ represents best-basis of the wavelet analysis QWPTBB, $\{\psi_g, \psi_h\}$ represent the pair of mother wavelet, and P is energy of wavelet coefficients. As QWT and QWPTBB use dual-tree approaches or two different wavelet filter-banks ,$\{\psi_g, \psi_h\}$, which are designed to form analytical complex filter banks, $\{\psi_g(x, y) + j\psi_h(x, y), \psi_h(x, y) \approx H\left(\psi_g(x, y)\right)\}$. The magnitudes of the wavelet coefficients are proven to be shift-invariant [137]; therefore, their corresponding wavelet-based sub-band descriptors, e_{QWT} and e_{QWPTBB}, are as well shift-invariant. Thus, QWT and QWPTBB wavelet would provide better descriptors than their real counterparts DWT and DWPTBB according to five criteria of Starck [120].

4.3.5 Information Measurement

In Sub-Section 4.3.4, four different wavelet transforms generate corresponding wavelet-based sub-band descriptors, $e = \{e_{DWT}, e_{DWPTBB}, e_{QWT}, e_{QWPTBB}\}$. With these descriptors e, energy probability distribution density p_e of PDF at each scale s_j can be computed as follows

$$p_e(x,y,s_j) \triangleq p\{\mathbf{P}\left[\mathbf{w_i} f(x,y,s_j)\right] |_{j<=m}\} \tag{4.3.50}$$

$$= \frac{\mathbf{P}\left[\mathbf{w_i} f(x,y,s_j)\right]}{\sum_j \sum_i \mathbf{P}\left[\mathbf{w_i}, f(x,y,s_j)\right]} \tag{4.3.51}$$

The above formula computes the probability of energy density at one location (x,y) across different sub-bands from the smallest scale, 1, to currently considered scales, m. The first wavelet level analyses images with the smallest size of scanning windows; therefore, it generates coefficients from finest details of images. Then, the sampling window sizes are doubled after each level; they generate coarser details. It is quite similar to PSS sampling operations except that scales are doubled rather than increased by a unit. Like PDF of PSS descriptor, WSS descriptors PDF are distributed with increasing scales of j, from level 1 (smallest size of scanning windows) to level m (currently biggest size of scanning windows). From the equation 4.3.51, it is straightforward to compute feature-space entropy $H_E\,(x,y,s_m)$ as follows

$$H_E = -\sum_{j<=m} p_e(x,y,s_j) \log(p_e(x,y,s_j)) \tag{4.3.52}$$

where $p_e(x,y,s_j)$ is sub-band energy density of wavelet-based descriptors' PDF, defined in the Equation 4.3.51. It is noted that both entropy of PSS's descriptors and the proposed WSS descriptor only characterizes statistics of local spatial information with respect to PSS's a window size and WSS's scale levels. These descriptors lack of global information with respect to the whole image as well as the whole set of natural images [127].

Equations 4.3.51 and 4.3.52 allows computation of feature-space entropy, H_E, according to its definition in Equation 4.3.36 for WSS sub-band energy descriptors at each specific scale. The rest of computational problem lies in estimation of inter-scale saliency $W_D(s,\vec{x})$ with WSS descriptors; in other words, how its definition in Equation 4.3.37 should be interpreted computationally. In Equation 4.3.36, an inter-scale saliency is measured as total variation in probability distribution at two consecutive scales. In PSS, the same set of pixel-value descriptors, or pixel values, (d) is in PDF of different scales; therefore, PSS needs to reconstruct PDF according to its incremental window

sizes. One advantage of WSS over PSS is uniqueness of non-overlapping wavelet-based sub-band descriptors since wavelet analysis aims at sparse representation of images. In WSS, PDF at level $m+1$ can be constructed by concatenation of already known descriptor set M from all levels $\{1 \dots m\}$ and a descriptor set E from the current level $m+1$; then, the Equation 4.3.51 is utilized on the concatenated set of descriptors (M, E) for computation of sub-band energy density. Though this characteristic simplifies construction of PDF across wavelet levels, it complicates formulation of their differences, as defined in Equation 4.3.37. Since two different sets of descriptors make total variation of two PDF unjustifiable, an alternative interpretation of inter-scale saliency, W_E as defined in Equation 4.3.37, is needed. Let's consider $P(M) = \{p_e(x, y, s_j) | s_j <= m\}$, PDF of all sub-bands up to the current level, m. When a wavelet-based sub-band descriptor, e, at level $m+1$ are generated, the additional sub-band descriptor will alter the overall PDF. The difference between the prior PDF, $P(M)$ and the modified PDF, $P(M|e)$, can be measured by Kullback-Leibler divergence as follows.

$$K(P(M|e), P(M)) = \int_M P(M|e) \log \frac{P(M|e)}{P(M)} \tag{4.3.53}$$

It is noted that similar concept is also utilized in Itti's Bayesian Surprise Saliency (BSS) metric [93]. Moreover, the measurement in the Equation 4.3.53 can be extended for multiple sub-bands descriptors. In other words, integration of the KLD, as shown in Equation 4.3.53, over multiple additional sub-bands descriptors $E = \{e_1 \dots e_i\}$ becomes mutual information between the current set of descriptors M and the additional set of descriptors E. In other words, an expectation of surprise or inter-scale saliency, W_E, for integrating the set of descriptors E of the level $m+1$ into the current set of descriptors M at the level m is measured by their mutual information, shown in Equation 4.3.54.

$$W_E \triangleq MI(E, M) = \int_{e \in E} K(P(M|e), P(M)) = \int_{E,M} P(E, M) \log \frac{P(E, M)}{P(M)P(E)} \tag{4.3.54}$$

The mutual information is chosen as inter-scale saliency, W_E, since it measures differences in amount of information from two consecutive scales. In other words, Equation 4.3.54 realizes the computation of inter-scale saliency W_E, as defined by Equation 4.3.37, for wavelet-based sub-band descriptors. Furthermore, usage of mutual information as inter-scale saliency measurement well emphasizes the structural coherence of image data across scales. If there are useful structures such as edges or joints in images and these structures would be consistent across multiple consecutive scales. Hence, it generates mutual information between data of two scales. Zero mutual information means noisy image data across scales since noises have no correlation across scales as

its self mutual-information is zero, $I(N,N) = 0$. Therefore, computation of inter-scale saliency W_E as mutual information satisfies the fifth criterion of the good estimation by Starck *et al.*[120]. The mutual information can be estimated by Equation 4.3.54 if PDF is continuous. However, wavelet-based descriptors are discrete; therefore, an alternative formula for mutual information of descrete signals [79] is used in Equation 4.3.55. The mutual information between discrete sets (E,M) is difference of their total entropy before concatenation, $(H(E) + H(M))$, and after concatenation, $H(E,M)$. Three entropies $H(M)$, $H(E)$ and $H(E,M)$ are calculated from wavelet-based sub-band descriptors by Equations 4.3.56, 4.3.57 and 4.3.58.

$$W_E \triangleq MI(E,M) = H(E) + H(M) - H(E,M) \tag{4.3.55}$$

$$H(M) = -\sum_{s_j \leq m} p_e(x,y,s_j)\log(p_e(x,y,s_j)) \tag{4.3.56}$$

$$H(E) = -\sum_{s_j = m} p_e(x,y,s_j)\log(p_e(x,y,s_j)) \tag{4.3.57}$$

$$H(E,M) = -\sum_{s_j \leq m+1} p_e(x,y,s_j)\log(p_e(x,y,s_j)) \tag{4.3.58}$$

Where W_E is the inter-scale saliency, $MI(E,M)$ is computed as differences between total entropy $(H(E) + H(M))$ and joint entropy $(H(E,M))$, p_e is sub-band energy density at locations (x,y) and scale s_j. It is noted that the joint entropy $H(E,M)$ can be reused as $H(M)$ for the next level inter-scale saliency estimation due to uniqueness of the sub-band descriptors.

From only definitions, as shown in Equations 4.3.35-4.3.39, of WSS and DIS, scale-saliency for wavelet-based descriptors, this Sub-Section has described in details how each of those definitions can be calculated with descriptors e and its energy density p_e. All formula for computation of WSS saliency are written as follows.

$$Y(M_{s_p}) = H(M_{s_p}) * W(E_{s_p}) \tag{4.3.59}$$

$$H(M_{s_p}) = -\sum_{s_j \leq s_p} p_e(s_j)\log(p_e(s_j)) \tag{4.3.60}$$

$$W(E_{s_p}) \triangleq MI(E_{s_p}, M_{s_{p-1}})) = H(E_{s_p}) + H(M_{s_p-1}) - H(E_{s_p}, M_{s_p-1}) \tag{4.3.61}$$

$$\vec{s_p} \triangleq \{s : H(M_{s-1}) < H(M_s) \wedge H(M_s) > H(M_{s+1})\} \tag{4.3.62}$$

Where $Y(M_{s_p})$ and (M_{s_p}) is scale saliency and entropy of the descriptor set M_{s_p} at level s_p and they are computed by Equations 4.3.59 and 4.3.60 accordingly. Inter-scale saliency $W(E_{s_p})$ is computed from mutual information between a descriptor set E_{s_p}, the current descriptor at level s_p, and a descriptor set M_{s_p-1}, descriptors of all levels

$s_1 \dots s_p$. In WSS, the characteristic scale s_p is chosen to maximize feature-space information $H(M_s)$. This selection of the characteristic scale s_p according to maximization of feature-space entropy $H(M_s)$ is adapted from original proposal of saliency with PSS descriptors [22]. However, this selection strategy is error-prone especially to images with complex small or random textures or details since WSS easily chooses s_p from small scales containing unnecessary details and textures while neglects bigger scales containing useful structures of images. With bias of Shannon entropy toward noise, WSS might fail to enclose any useful structure.

This disadvantage of WSS can be avoided by modifying criteria for selecting characteristic scales s_p. The modification leads to a derivative of WSS called Discriminant Information Scale (DIS) selection which selects s_p in order to maximize inter-scale saliency, $W(E_{s_p})$ instead. In this chapter, DIS refers to a scale selection method instead of Gao's saliency method [108]. The only difference between DIS and WSS is how the characteristic scale s_p is found in Equation 4.3.62 and it is re-written for DIS as follows.

$$\vec{s_p} \triangleq \{s : W(E_{s-1}) < W(E_s) \wedge W(E_s) > W(E_{s+1})\} \tag{4.3.63}$$

Where $W(E_{s-1})$, $W(E_{s-1})$, and $W(E_{s+1})$ are inter-scale saliency of the set descriptor E at levels $s-1$, s, and $s+1$. Experiments with DIS and WSS are carried out and results are detailed in the next section in order to confirm effectiveness of the proposed strategies.

4.4 DISCUSSION & RESULTS

The previous Sections 4.3.4 and 4.3.5 have analysed theoretical advantages of **WSS** and its derivative **DIS**. In addition, the Sub-Section 4.3.4 presents four descriptors based on different wavelet transforms: DWT, QWT, DWPTBB, and QWPTBB. Accordingly, there are eight derivatives for the proposed method according to specific choices of scale selection mechanisms and sub-band descriptors, for example, DWT-WSS,QWT-WSS,DWPTBB-WSS, etc. Putting all these derivatives into one big experiment probably causes confusion. ;therefore, several small experiments are carried in the following manner. For each database, there are two separate sets of experiments and results for WSS and DIS. At either experiment for WSS or DIS, four types descriptors DWT, QWT, DWPTBB, and QWPTBB are utilised. In this section, experiments are carried out on Niel Bruce's and Kootra's databases; therefore, experimental results, tables and graphs, are presented in following structure.

- Neil Bruce's Database - WSS

Table 4.2: Quantitative Result of Neil Bruce's Database- WSS

Figure 4.8: ROC Curve of Neil Bruce's Database - WSS

- Neil Bruce's Database - DIS

  Table 4.3: Quantitative Result of Neil Bruce's Database - DIS

  Figure 4.9: ROC Curve of Neil Bruce's Database - DIS

- Kootstra's Database - WSS

  Table 4.4: Quantitative Result of Kootstra's Database - WSS

  Figure 4.10: ROC Curve of Kootstra's Database - WSS

- Kootstra's Database - WSS

  Table 4.5: Quantitative Result of Kootstra's Database - DIS

  Figure 4.11: ROC Curve of Kootstra's Database - DIS

To evaluate the proposed methods against other saliency approaches, they are compared with AIM model [4], PSS [22], and BSS [19]. The purpose of comparisons are not for claiming the best saliency method or the highest possible evaluating measurement but to prove that the rationale of the assumption that feature and structural complexity would be good clues for human attention . Results of simulations much depends on several factors such as choices of databases, elimination of experimental bias, performance of human test subjects, etc. In evaluation of saliency methods, different researchers choose different datasets and measurement methods or even create their own database. In this research, the main target is evaluating the effectiveness of information measurement in visual attention; therefore, the most common processes and databases would be chosen to confirm generalization of the assumption.

Bruce and Tsotsos [71] database is certainly among the popularly used stimuli to evaluate informative clues. However, only Bruce's database is certainly not sufficient enough due to limits in numbers and contents of stimuli. Therefore, Kootstra's database [119] is chosen for additional testing samples and ground-truths. Two databases with over 200 samples and ground-truths provided by more than 50 human subjects would help to confirm the generalization of the proposed frameworks to a certain extent. Similarly, only common evaluation approaches are deployed in this study, and including both quantitative or qualitative methods. Quantitative relations between different saliency methods and human visual performance are shown by appropriate statistical methods (AUC,NSS) with eye-tracking data as ground-truths. Meanwhile, the qualitative



Figure 4.6: Neil Bruce's database

results, visual comparisons of different saliency maps, gives a glimpse about performance for each individual sample. It also specifies imaging contexts where saliency methods give reasonable solution as well as situations where saliency maps are unreasonable to human perception.

4.4.1 Databases of image stimuli

The ground-truth and data for basic evaluations of visual saliency performance is obtained from eye-tracking experiments. Specifically in Neil Bruce database, 120 different color images are observed in random orders while there are 4 seconds gap between the previous and the next stimuli. To ensure consistency and accuracy of the database, subjects are asked to seat 0.75 m in front of a 21 inch CRT monitor. Human subjects receive no further instructions for any actions or clues for what images appear next. Furthermore, image contents are varied from indoor to outdoor environments. Sometimes, there are clear interesting objects in the scene; while some scenes are really general without any particular interests in any subjects. A non-head mount eye tracking apparatus extracts locations of eye-fixation while human test subjects look at sample images. Other setting-up parameters are intended for a general-scene based stimuli which are typically found in urban environments. Moreover, the same parameters are used for collecting data from 20 different subject over 120 testing samples. The following Figure 4.6 shows first eight images from the Neil Bruce's database. Despite its popularity, images from Neil Bruce's database has narrow semantic content since it contains only urban scenes and mainly indoor environments. Besides that, the number of samples are relatively small; hence, an additional database should be included in simulations such that there are more testing images of natural objects like animals, flowers, in natural environments. Kootstra's database [119] have just satisfied these requirements with additional ground-truths for further experiments; moreover, the ground-truths



Figure 4.7: Kootstra's database

data are also collected from eye-tracking experiments although the experimental process is slightly different from what have been done to collect Bruce's database. In those psychological experiments, with head-mount eye-tracking devices, thirty-one students (15 men, 16 womens) ranging from 17 to 32 of age took part in the experiments, and they are all naive about aims of experiments. Each human subject observes a total of 99 photographic image in five different categories while their eye movements are recorded simultaneously with the head-mounted device. There are nine-teen images in natural symmetry category; each of which contains symmetrical natural objects. Beside such sceces, other non-symmetrical photographic scenes are as well included into the image databases such as: 12 images of animals in natural habitats, 12 images of street environments, 12 images of street scenes, 16 images of building and 40 images of natural environments. Figure 4.7 gives an example of 5 image categories in the Kootstra's database. Noted that, each image is presented to viewers with a resolution of 1024x768 pixels on an 18" CRT monitor at a distance of 70 cm from the participant.

4.4.2 Quantitative Comparisons of Saliency Methods

The quantitative performance includes Receiver Operating Characteristics (ROC) curves with Area Under ROC Curve (AUC), and Normalized Scanpath Saliency (NSS) as numerical results. To ensure fair comparisons between methods, open-source evaluation codes for LCC, NSS and AUC [138] are employed. Saliency maps are stan-

dardized around median instead mean of distributions. Quantitative evaluation of visual saliency map on natural images with eye-tracking data ground has been initially studied by Tatler and recently summarized by Borji *et al.*[138]. More information about mechanisms behind ROC and AUC can be found [138]. This section focuses on analysing quantitative results to compare, evaluate and prove rationale of the proposed approach. As the main purpose of this evaluation is confirming effectiveness of informative clues in human's visual attention, the framework is not optimally tuned to reach the maximum AUC or NSS by addition of blur and central bias [112].

All four descriptors mentioned in the Section 4.3.4 have been simulated with image samples from both Neil Bruce's and Kootstra's datasets. Noted that scale selection mechanisms might have influences in formation of saliency maps; therefore, two separated simulations are carried out to investigate that effect as well. Figure 4.8 and 4.9 summarize simulations' results of the proposed methods with corresponding WSS and DIS respectively in Neil Bruce's image dataset.

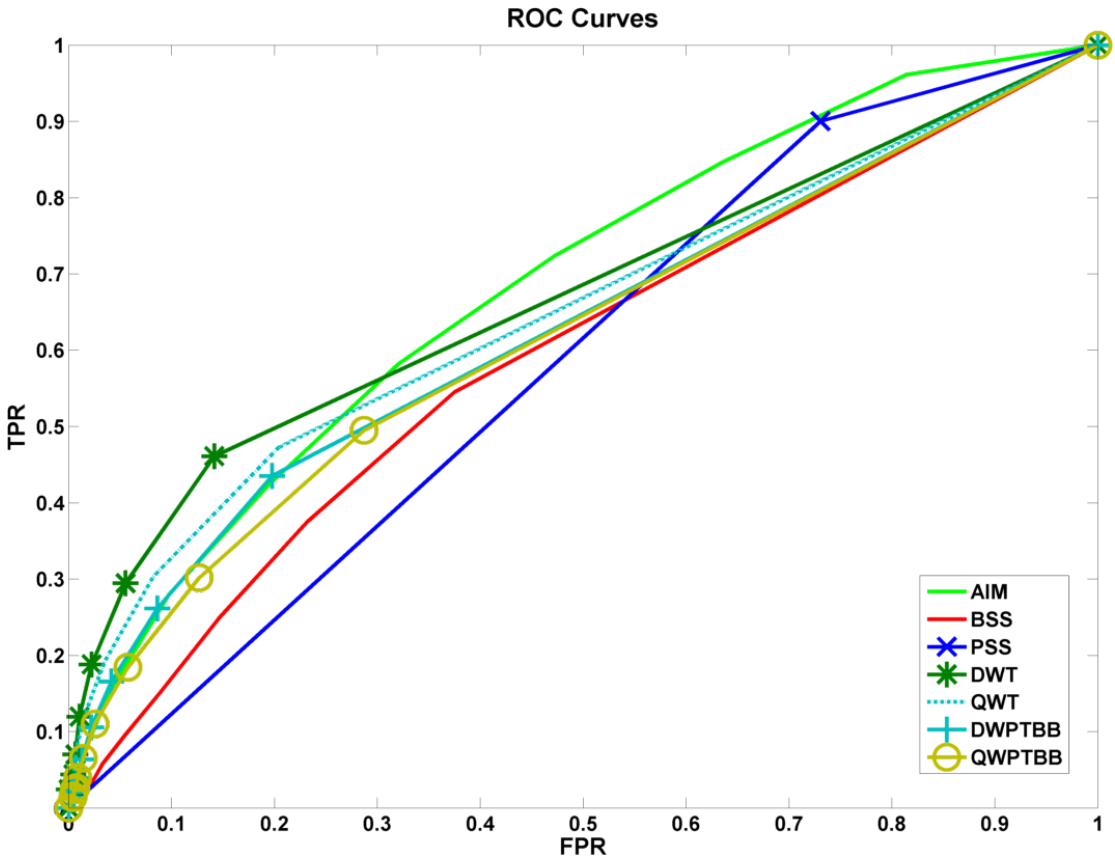


Figure 4.8: ROC Curve of Neil Bruce's Database - WSS

Table 4.2: Quantitative Result of Neil Bruce's Database- WSS

MTH	LCC	NSS	AUC	TIME(s)
AIM	0.1576	0.12378	0.72382	50.41714
BSS	0.04823	0.54031	0.59728	5.37500
PSS	0.07685	-0.39175	0.58558	7.10922
DWT	-0.01275	0.33358	0.67822	1.2401
QWT	-0.02374	0.30002	0.66263	1.92310
DWPTBB	-0.03596	0.26079	0.64144	2.21160
QWPTBB	-0.03992	0.23714	0.63531	5.28395

As depicted in Figure 4.8 and Table 4.2, performances of four WSS derivatives in all three numerical evaluation LCC, NSS and AUC are decreased in the following order DWT, QWT, DWPTBB and QWPTBB. It is noted that each method has its particular colour, described by legends in the Figure 4.8, for its ROC curves and numerical results in each row of the Table 4.2. The best basis approaches *DWPTBB* and *QWPTBB* produce poorer results than *DWT* and *QWT* in all four quantitative measurement LCC, NSS, AUC, and TIME. In AUC evalutation, all WSSs are better than PSS and BSS but worse than AIM which means WSSs' saliency maps predict eye-fixation points better than PSS and BSS but worse than AIM. In NSS evaluation, all WSSs performs better than PSS and AIM but worse than BSS which means responses of WSSs saliency

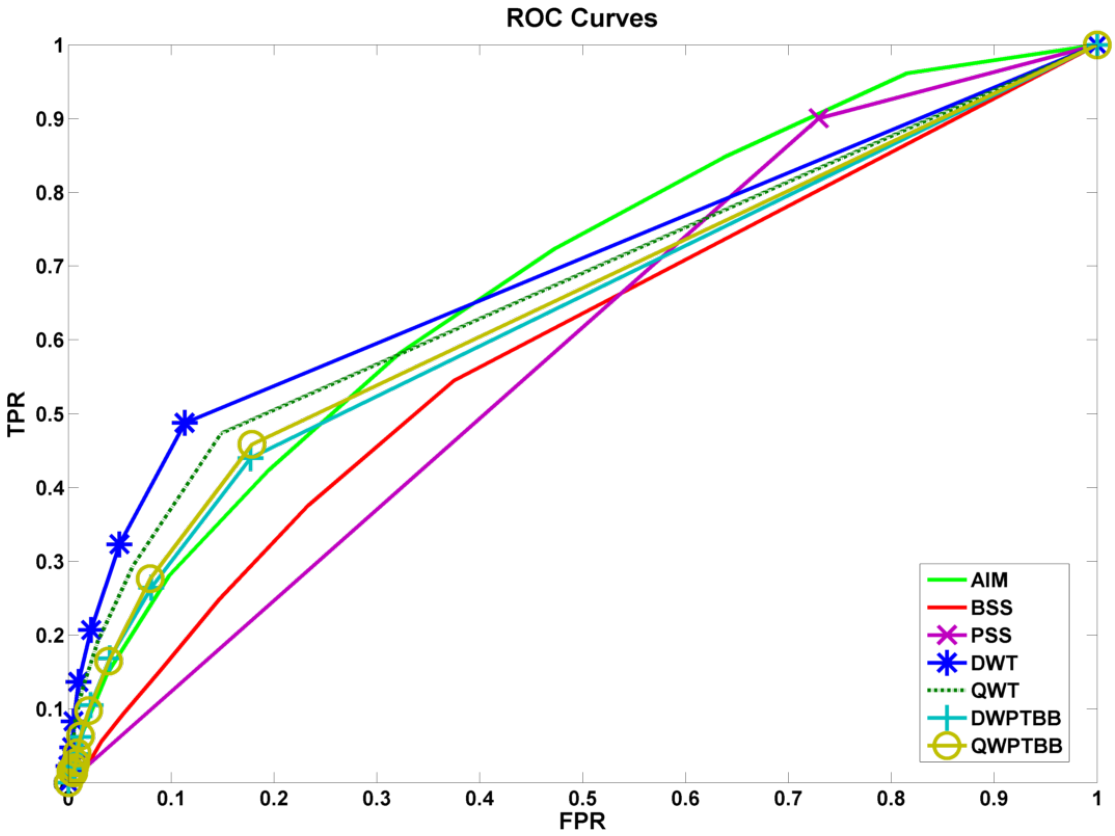


Figure 4.9: ROC Curve of Neil Bruce's database - DIS

Table 4.3: Quantitative Result of Neil Bruce's database - DIS

Methods	LCC	NSS	AUC	TIME(s)
AIM	0.1576	0.12378	0.72382	50.41714
BSS	0.04823	0.54031	0.59728	5.37500
PSS	0.07685	-0.39175	0.58558	7.10922
DWT	-0.01731	0.31782	0.70288	1.26889
QWT	-0.02482	0.29877	0.68286	1.88938
DWPTBB	-0.03314	0.27559	0.65201	2.21160
QWPTBB	-0.04513	0.23941	0.66066	5.36596

values at saccade points are better than those of AIM and PSS but worse than BSS. Meanwhile, all WSSs, in LCC evaluation, are worse than all reference method AIM, BSS ad PSS which means saliency maps of AIM, BSS and PSS linearly correlate with ground-truth saliency maps better than WSSs. The maximum computational time of a WSS derivative per image is less than that of PSS. It is noted that the PSS is implemented in C++ with MATLAB interface and entire WSS codes are written in MATLAB. It proves that WSSs are more efficient scale-saliency than PSS is. In comparison with other information-based saliency approaches AIM (MATLAB code) and BSS (C Code), WSSs (MATLAB code) is comparable to BSS and much faster than AIM.

Quantitative performances of four **DIS** methods are shown in Figure 4.9 and Table 4.3

in which each methods is presented with its own colour according to legends in the Figure 4.9. When compared with AIM, BSS and PSS methods in all four quantities, performances of DIS and WSS are similar. When compared with each other, AUC of DWT and QWT descriptors with DIS approaches are better than AUC of WSS derivatives. However, performances of DIS and WSS are equivalent in three other evaluation quantities.

Above is shown simulation results from Neil Bruce image data-set with eye-tracking locations. Despite of its recently popular database in evaluating saliency maps, the data-set has limitations analysed in the subsection 4.4.1. Another set of images should be brought in to enhance diversity of testing samples. Kootstra's database with more image categories and all eye-tracking data ground truth is a perfect candidate. Additional simulation results would confirm and generalize rationale of the proposed information-based saliency methods. Similar to the Table 4.2 and Figure 4.8, Table 4.4 and Figure 4.10 demonstrate how well the proposed methods with four descriptors and WSS scale selection mechanism perform against information-based saliency methods like BSS, AIM and pixel-based scale saliency PSS. It is noted that colour schemes of Tables 4.2 4.3 and Figures 4.8 4.8 are also employed.

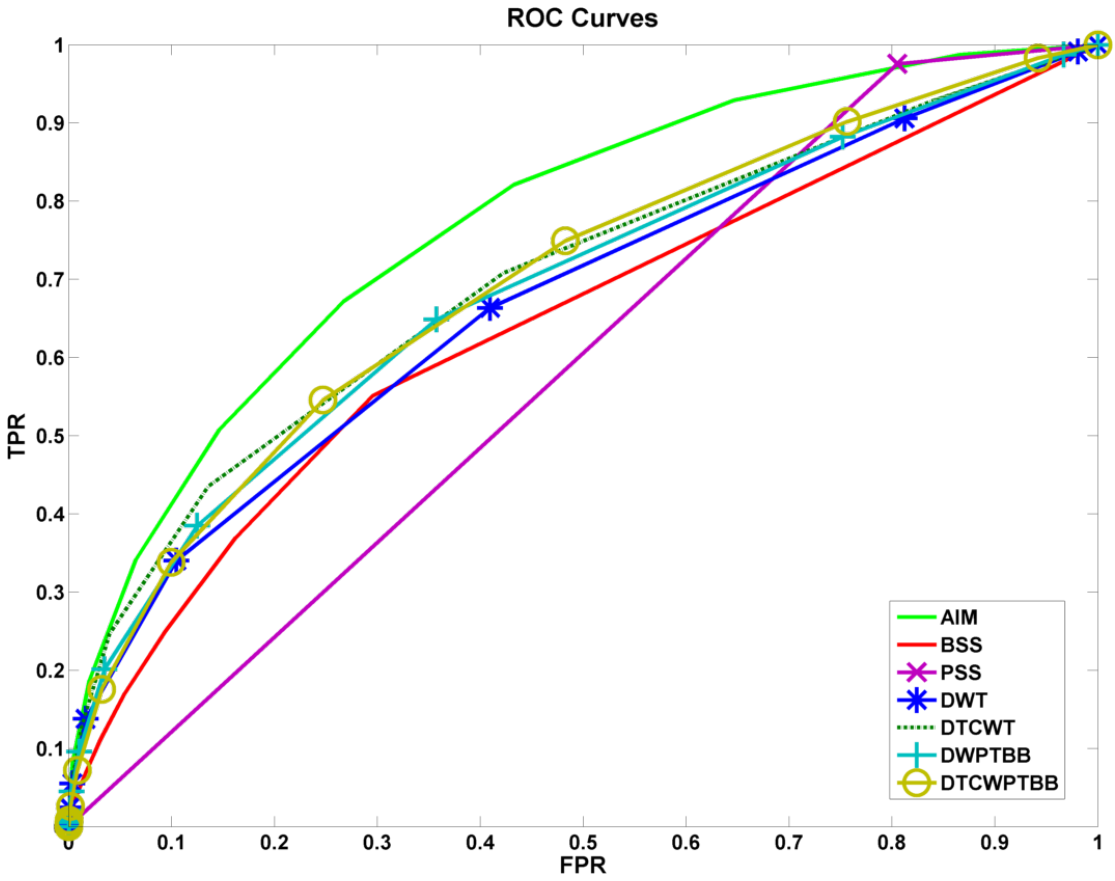


Figure 4.10: ROC Curve of Kootstra's database - WSS

Table 4.4: Quantitative Result of Kootstra's database - WSS

Methods	LCC	NSS	AUC	TIME(s)
AIM	0.20247	0.47584	0.79975	128.75095
BSS	0.17726	0.71157	0.64577	27.17000
PSS	0.12626	-0.35322	0.58492	17.06625
DWT	0.11896	0.48489	0.71497	0.49332
QWT	0.13280	0.50702	0.72980	1.33988
DWPTBB	0.10858	0.45557	0.71732	5.12908
QWPTBB	0.08145	0.34678	0.73817	3.98680

In evaluation results of Kootstra Database [23], the best result in terms of AUC is from QWPTBB descriptor among all four descriptors, and the second best is QWT; while both DWT and DWPTBB have nearly equal AUC values. Compared with AIM, BSS, and PSS, AUC values of WSS derivatives are larger than those of BSS and PSS but smaller than that of AIM. It means WSSs predict eye-fixation better than BSS, PSS do but worse than AIM does. Meanwhile, NSS values of WSS derivatives are larger than PSS, equal to AIM and less thang AIM. As NSS measures how well saliency maps response to saccade, saliency responses at WSSs' saliency maps of saccade points are better than PSS, equal to AIM and less than BSS. Furthermore, LCC values of WSS

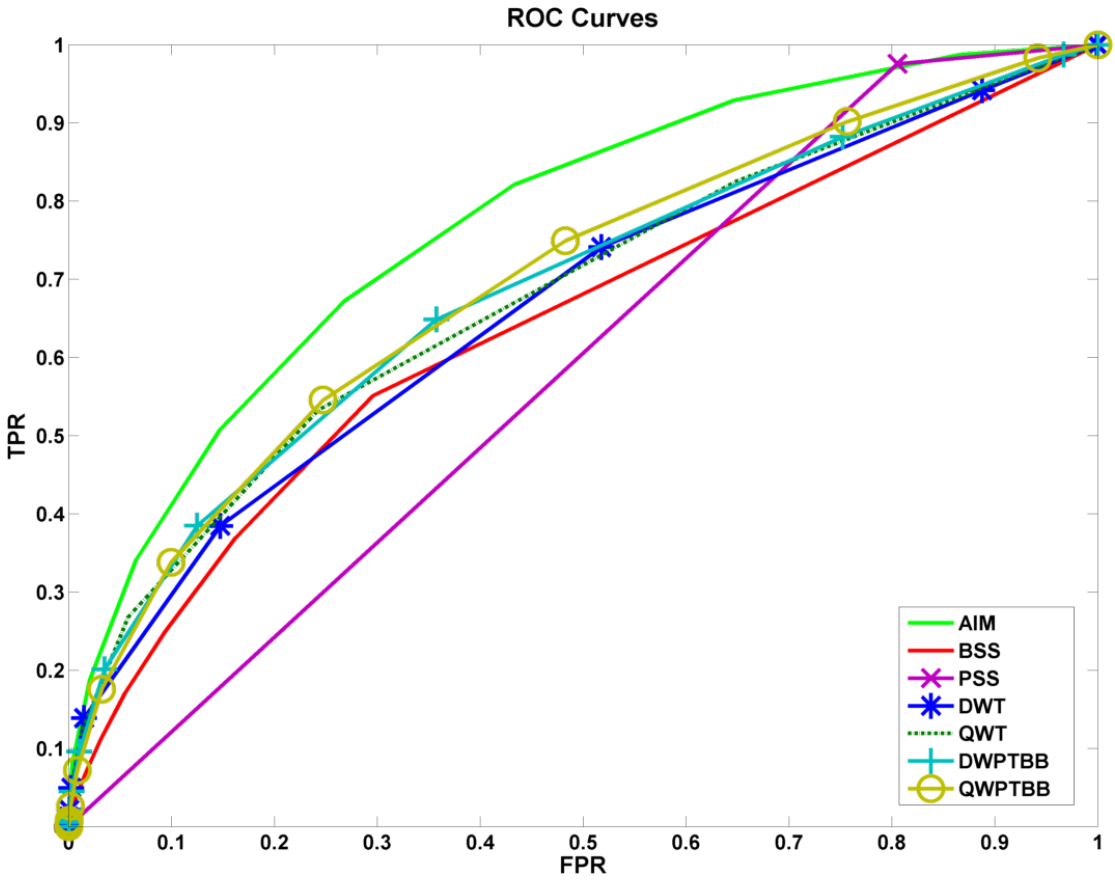


Figure 4.11: ROC Curve of Kootstra's database - DIS

Table 4.5: Quantitative Result of Kootstra's database - DIS

Methods	LCC	NSS	AUC	TIME(s)
AIM	0.20247	0.47584	0.79975	128.75095
BSS	0.17726	0.71157	0.64577	27.17000
PSS	0.12626	-0.35322	0.58492	17.06625
DWT	0.12006	0.48471	0.70598	0.45579
QWT	0.12493	0.50270	0.71735	2.68730
DWPTBB	0.10858	0.45557	0.71724	4.55148
QWPTBB	0.08145	0.34678	0.73785	4.45500

derivatives are equal to PSS but less than BSS and AIM, which means linear correlation between WSS's saliency maps and ground-truth saliency maps is as much as this correlation of PSS saliency maps but worse than those of BSS and AIM. Though WSSs do not give the best results any numerical evaluation, this result shows comparable performances of WSS against other information-based saliency maps like BSS and AIM. Moreover, as WSSs are better than PSS in numerical evaluations, sub-band wavelet descriptors are better descriptors for scale-saliency computation. Furthermore, WSSs take the least computational time to process a frame among all considered saliency maps, which shows its efficiency and fast information estimation. Besides WSS scale

selection, experiments are done to compare how DIS scale selection performs on same databases with suggested sub-band descriptors in Figure 4.11 and Table 4.5 with particular colours for particular methods. Almost all performances of DIS and WSS derivatives are similar though three are slight increment of AUCs and decrement of LCC and NSS in cases of DWT and QWT.

In general, the proposed methods performs quite well against other methods like BSS and AIM; especially, they are much better than PSS all evaluation criteria in both databases. Therefore, wavelet-subband energy descriptors are better than original pixel descriptors for constructing scale-saliency. DIS and WSS have similar performances in both Bruce's [4] and Kootstra's databases [23]. Therefore, choices of mechanisms for scale selections are irrelevant to general performances of the proposed saliency methods.

4.4.3 Qualitative Comparison of Saliency Methods

In this section, qualitative comparisons are carried out by visualizing a few examples of saliency maps from two mentioned database of Bruce and Kootstra. From each of the databases, only four test images are chosen to be displayed due to limited spaces though saliency maps are generated for every single image in either of the databases. The samples are intentionally chosen to show variety of contexts and scenes as well as they cover cases of successfully highlighting interested objects and cases of failing to emphasize salient regions. Along with the proposed methods, saliency-maps of AIM, BSS and PSS methods are also included in order to give visual comparisons to the proposed saliency methods. Directly below are displayed four samples from Neil Bruce's database.

Figures 4.12, 4.13, 4.14, 4.15 show saliency maps from four chosen gray-scale images, as noted as GRY. Generally, PSS identifies a large portion of images as salient regions (white regions), it explains why its average AUC, NSS in the Table 4.2 are the lowest, and BSS and AIM model give decent saliency maps for three over four samples. Four samples of saliency maps are deliberately chosen to show that different ranking of WSS, DIS derivatives, and their dependence on wavelet basis. Comparison between second row and third row of the Figures 4.12 to 4.15 shows that DIS provides sharper saliency maps with clearer small details, while WSS generates blurred saliency maps. Utilization of QWT and DWT basis generate quite similar saliency maps, as shown in the Figure 4.12. However, QWT-WSS performs better than DWT-WSS in Figure 4.13. DWPTBB-WSS and QWPTBB-WSS have their own advantages, especially in the case textured background, as shown in Figure 4.14. Finally, sometimes none of proposed methods gives reasonable saliency maps, Figure 4.15. It usually happens if images

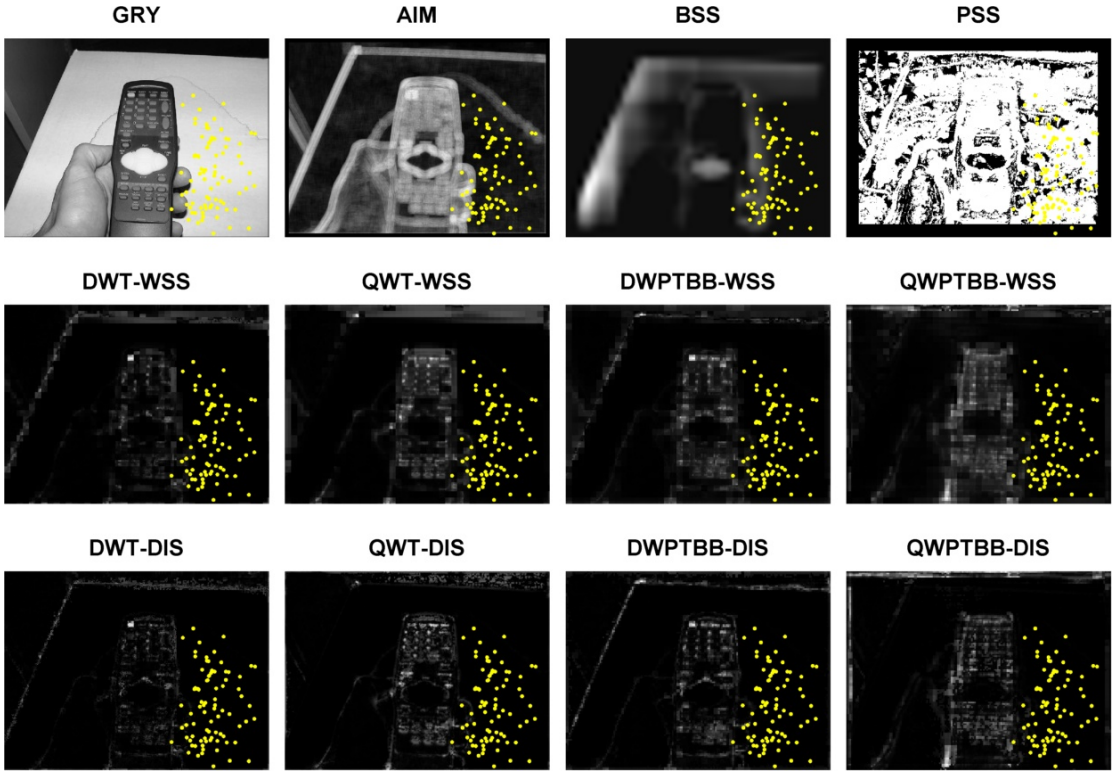


Figure 4.12: Saliency Maps 1

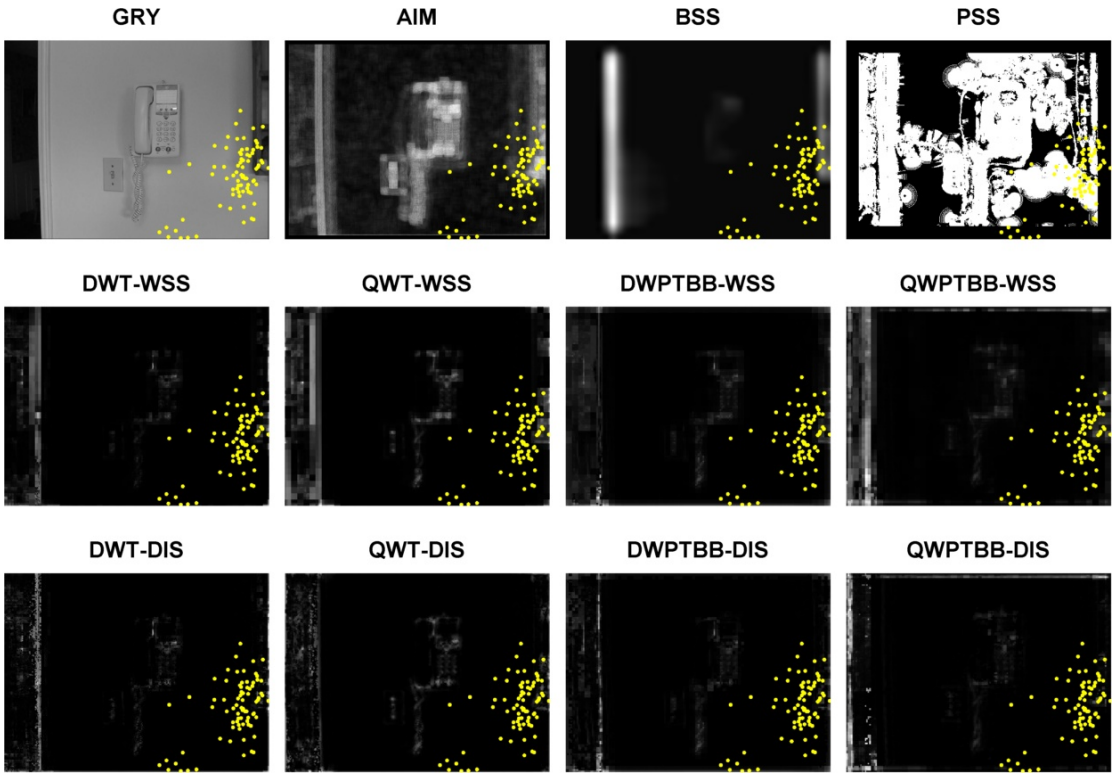


Figure 4.13: Saliency Maps 2

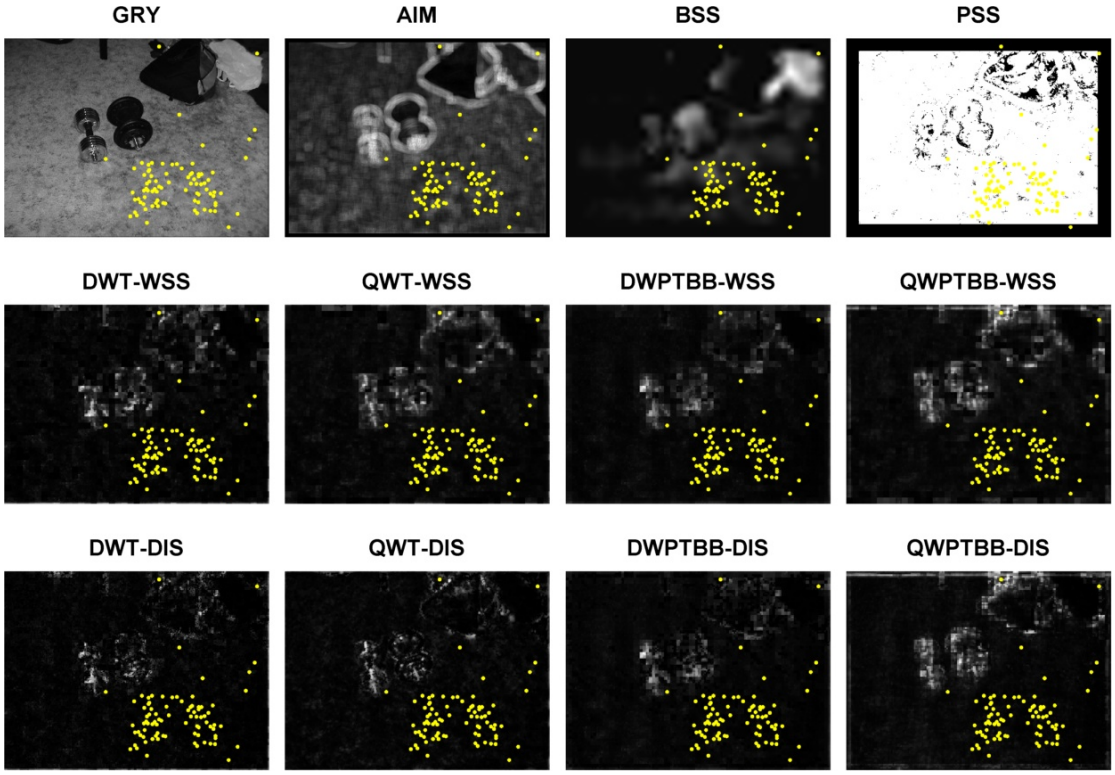


Figure 4.14: Saliency Map 3

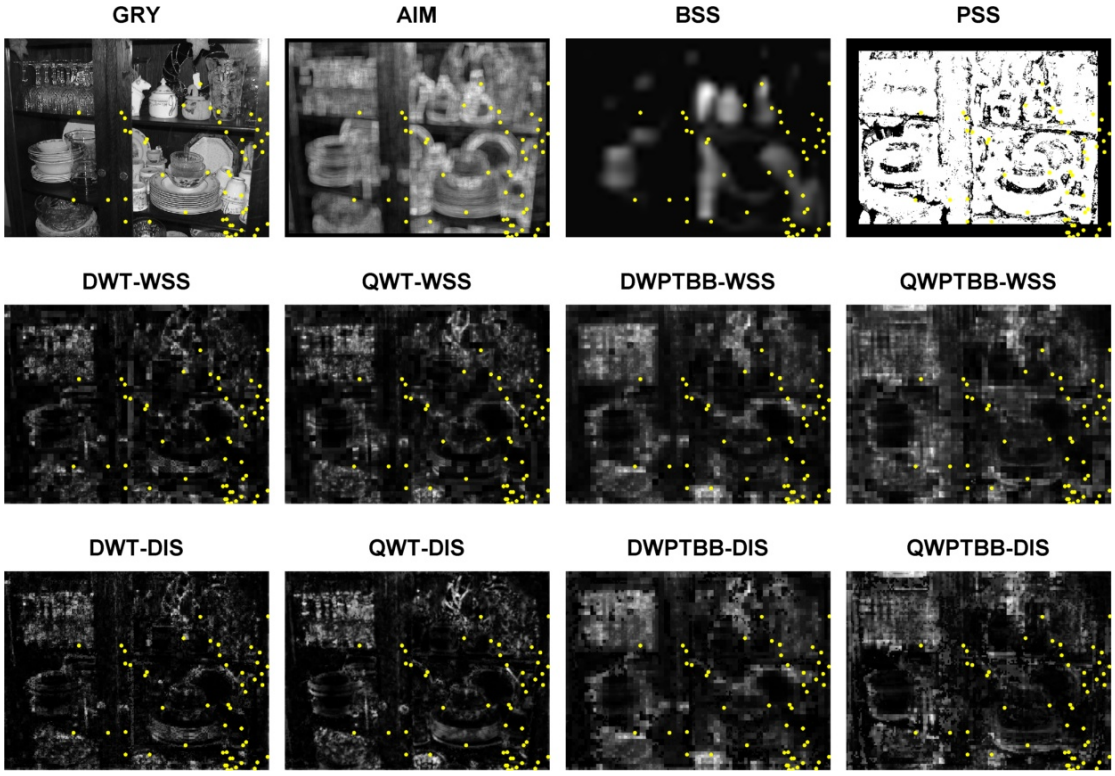


Figure 4.15: Saliency Map 4

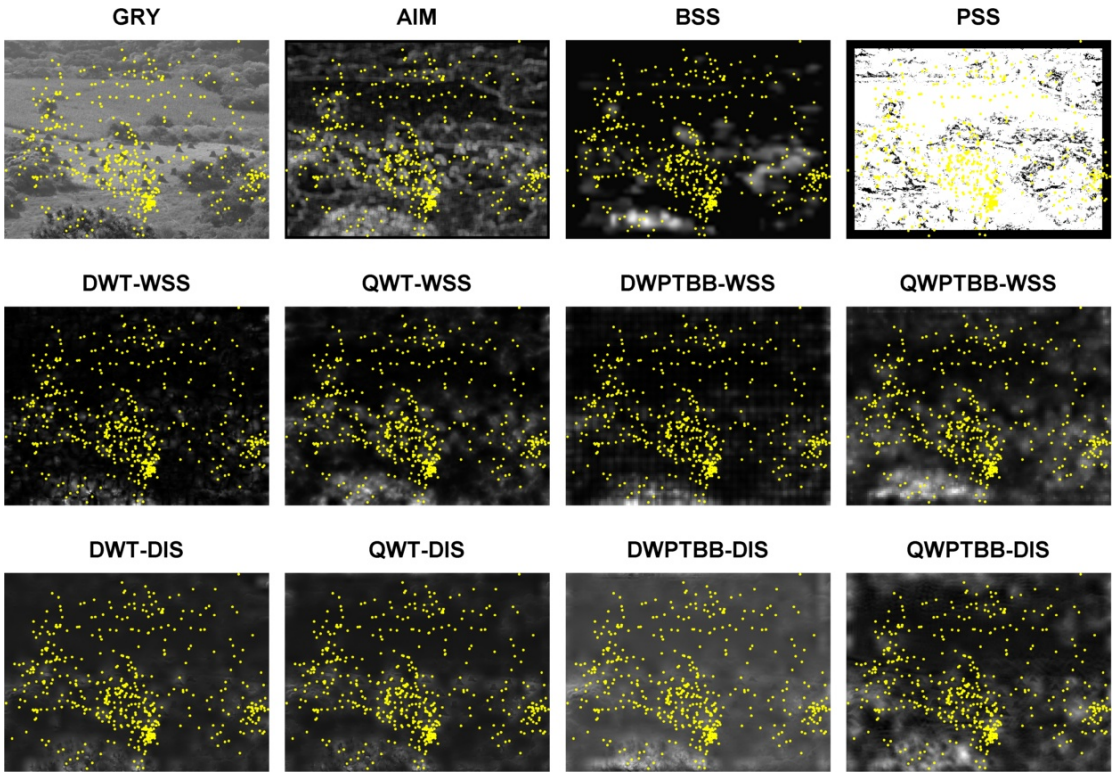


Figure 4.16: Saliency Map 5

are flooded with complex features. In all above samples, AIM saliency maps, the second column of the first row, provides a blurred, center-biased saliency maps, which are visually more appealing. Moreover, their distributions match to histogram of eye-fixations better; therefore, they tend to achieve better performance according the standard quantitative evaluation [112] . Blurring and histogram-matching saliency maps are popular optimization tricks to achieve better performance evalution. However, these techniques are all omitted in evaluating the proposed framework quantitatively and qualitatively so these experiments reveal true performance of wavelet-subband basis in scale saliency.

While Neil Bruce's data-set capture daily scenes in the urban and suburban areas, it lacks of ourscenes from natural landscapes. Therefore, its images do not represent the whole meaning of "natural images" category. In order to visually confirm effectiveness of the proposed methods, four "natural" samples with corresponding saliency maps from the Kootstra's database are shown in the following Figures 4.16, 4.17, 4.18 and 4.19. In accordance with gray-scale sample image, saliency maps of AIM, BSS, PSS and eight derivatives of the proposed wavelet scale saliency methods are displayed. The Figures 4.16 and 4.19 represent image of flowers taken in close range. Therefore, it shows quite a number of symmetric and small details. Meanwhile, Figures 4.17 and 4.18 contains the whole landscape of mountains and plateaus. Those scenes are usually

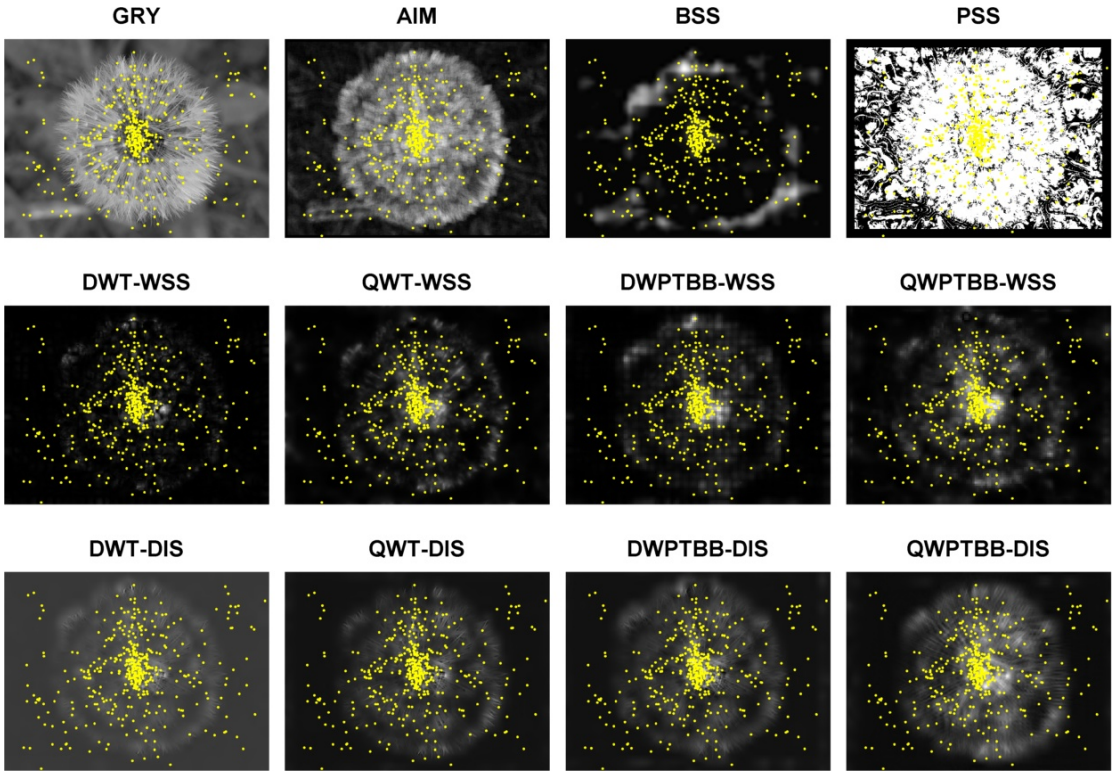


Figure 4.17: Saliency Map 6

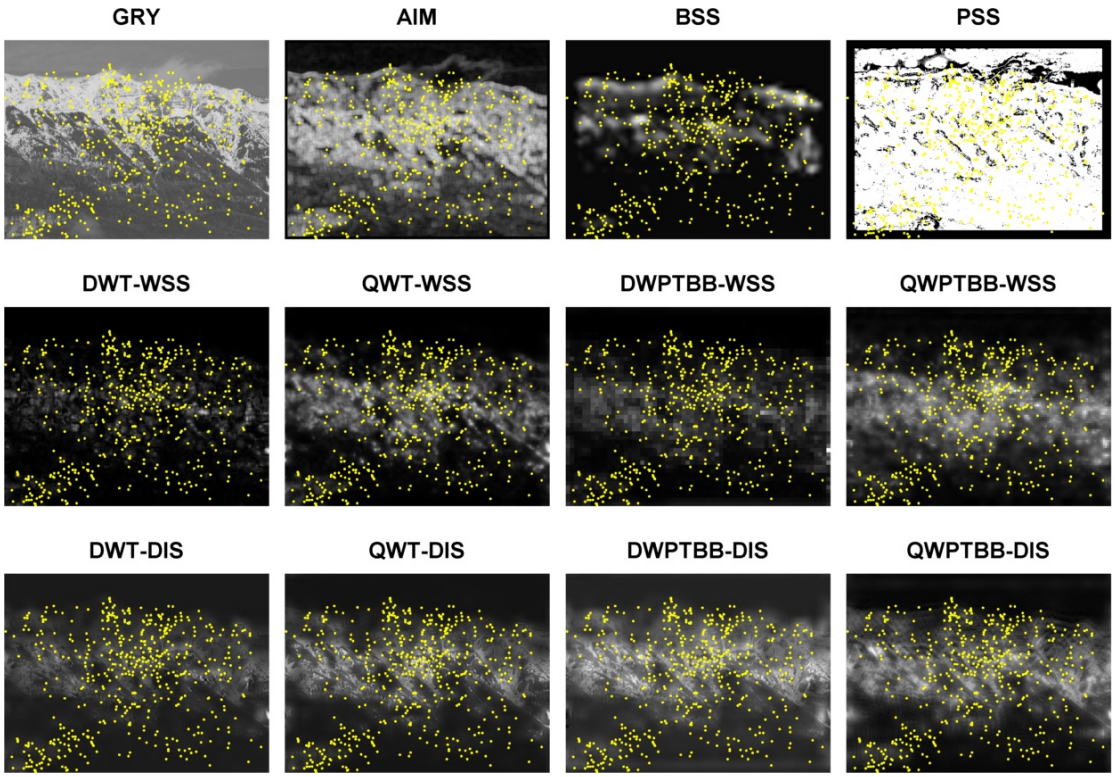


Figure 4.18: Saliency Map 7

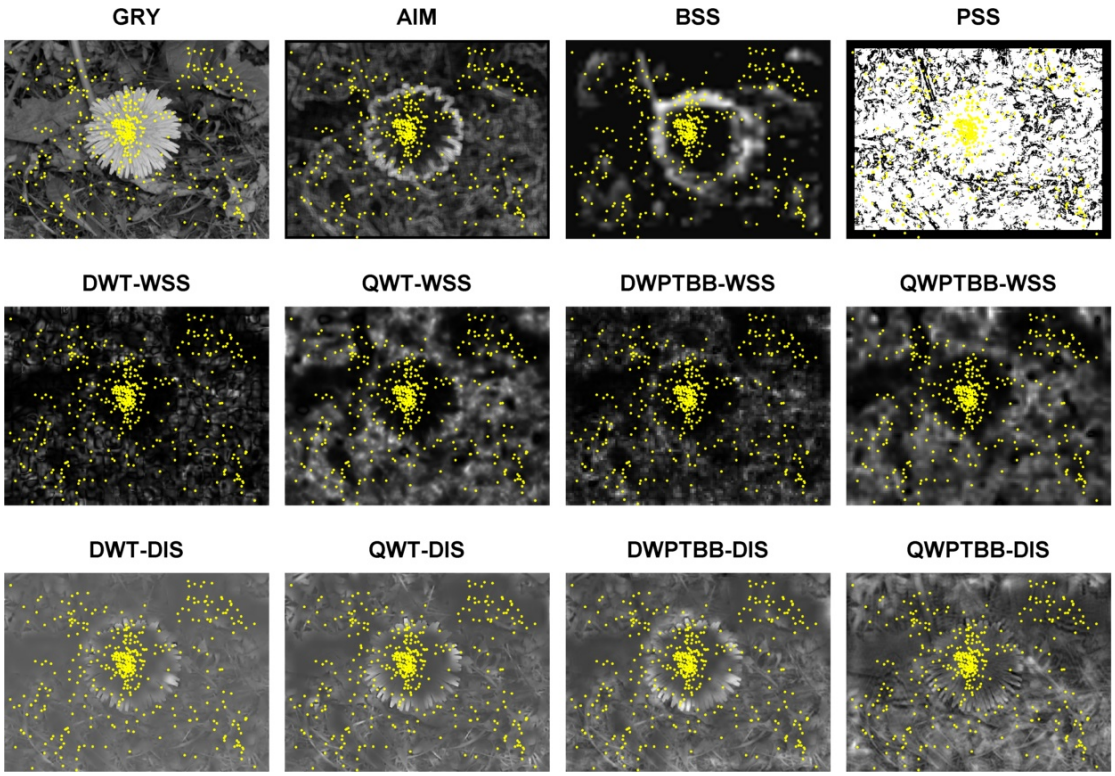


Figure 4.19: Saliency Map 8

anti-symmetric and much richer in information than flowery photos. In general observations, ITT method does the best job in selecting the salient features. Among derivatives of the proposed features, QWPTBB descriptors show the most competitive and comprehensive visual result, followed by QWT, DWTBB, and DWT based derivatives in descending order of performances. For comparison between WSS and DIS scale-selection mechanisms, there is slight but important difference between their saliency maps - WSS saliency maps in the second rows and DIS saliency maps in the third rows of the Figures 4.16,4.17,4.18, and 4.19. In these examples, WSS saliency maps tends to be more selective than DIS maps. In other words, WSS maps might have better discriminant power than DIS maps. It would explain why AUC and NSS results of WSS in the Table 4.4 are slight better than those of DIS in the Table 4.5. There are small changes in quantitatively visual results when different parameters are used. However, the proposed methods perform very well against other saliency methods like ITT and PSS.

4.5 CONCLUSION

In this chapter , an extension of scale saliency from pixel descriptors to sub-band energy density ones generated by four DWT, DWPTBB, QWT, and QWPTBB wavelet trans-

forms with two different scale selection mechanisms WSS and DIS have been proposed. Comparing to pixel-value descriptors (PSS), the proposed ones are much more sparse but biased toward morphological shapes of mother wavelet. Moreover, they are more robust to external factors; therefore, generation of saliency maps are less affected by translational shifts and other affine transformation. Furthermore, wavelet packet descriptors with best basis algorithms are also considered since several psychological experiments suggest sparse factors in human vision system. Along with new descriptors, innovative coherent information framework for wavelet scale saliency is proposed and strong relations with Bayesian Surprise Model [93] are emphasized. Beside solid theoretical development, the experimental results are that the proposed methods are comparable to BSS model and surpasses the original scale saliency model PSS but worse than AIM.

CHAPTER 5

Multi-scale Discriminant Saliency (MDIS)

Supposed saliency is a binary classification between centre and surround classes, saliency value is measured as their discriminant power. As the features are defined by sizes of chosen windows, a saliency value at each location is varied accordingly. This chapter proposes computing saliency as discriminant power in multiple dyadic scales of Wavelet Hidden Markov Tree (HMT), in which two consecutive dyadic scales provide surrounding and central features, organized in a quad-tree structure. Their discriminant power is estimated as maximum a posterior probability (MAP) by Expectation-Maximization (EM) iterations. Then, a final saliency value is the maximum discriminant power generated among these scales. Standard quantitative tools and qualitative assessments are used for evaluating the proposed multi-scale discriminant saliency (MDIS) against the well-know information based approach AIM on its image collection with eye-tracking data. Simulation results are presented and analysed to verify the validity of MDIS as well as point out its limitation for further research direction.

5.1 INTRODUCTION

Multi-scale nature has been emphasized as the implicit part of human visual attention, this significant aspect of visual attention is often overlooked in several saliency algorithms. Hidden behind the characteristics are statistical dependencies across different scales of images. Simple observations on any normal photography show that the nearer two patches are to each other, the more similar their contents are. Despite that an across-scale relation is easily spotted, curse of dimensionality makes it difficult to model and estimate such information efficiently. Entropy-based saliency estimation

approach (ENT) in Chapter 3 only estimates entropy from patches in one block size; therefore, its saliency values do not include across-scale statistics. A Scale Space element is integrated in Wavelet Scale Saliency (WSS) of Chapter 4; however, only simple dependency between scales is captured by estimating from histogram of wavelet-based descriptors. In order to reveal more complex inter-scale dependence, more sophisticated statistical estimation of information such Wavelet-based Hidden Markov Model [139] needs to be utilised. Moreover, a different interpretation of saliency concepts is necessary to accommodate a better estimation technique. Gao .et.al [108] suggests to formulate saliency as a decision-theoretic problem which optimally minimizes classification error for centre and surround features. Their framework, called Discriminant Information Saliency (DIS) in this thesis, provides a flexible concept to model complex dependencies between a finer scale and a coarser scale across multiple levels. However, Discriminant Information Saliency (DIS) approach [17] considers only one fixed-size window; hence, it may lead to inconsideration of significant attentive features in a scene. Therefore, DIS approach needs constituting under the multi-scale framework to form multi-scale discriminant saliency (MDIS) approach, which the main motivation and contribution of this chapter which are organized as follows. Section 5.2 reviews principles behind DIS [108] and focuses on its important assumption and limitation. Section 5.2.1 provides alternative interpretation of centre-surround operation in statistical manner; moreover, it also describes the operation in the light of "large variance" / "small variance" Gaussian distribution as well as "implicit" / "explicit" manifolds concepts. After that, MDIS approach is carefully elaborated in Section 5.3 with several relevant contents such as multiple dyadic windows for binary classification problem in Sub-Section 5.3.1, multi-scale statistical modelling of wavelet coefficients though Hidden Markov Tree (HMT) framework as well as learning of HMT parameters in Sub-Sections 5.3.2, 5.3.3, maximum likelihood (MLL) and maximum a posterior probability (MAP) computation of dyadic sub-squares in Sub-Sections 5.3.4, 5.3.5. Though construction and optimisation of wavelet hidden markov tree are inspired by work of Romberg .et at. [140] and Choi [141] on image segmentation, MDIS formulates the across-scale statistical characteristics as continuous saliency values in saliency maps instead of binary decision (foreground/background) in segmentation maps. MDIS steps are combined for final saliency map generation in Sub-Section 5.3.6. Quantitative and qualitative analysis of the proposal with different modes are discussed in Section 5.4; moreover, comparisons of the proposed MDIS and the well-known information-based saliency method AIM [4] simulation data are presented with several interesting conclusions. Finally, main contributions of this chapter as well as further research direction are stated in the conclusion Section 5.5.

5.2 DISCRIMINANT SALIENCY (DIS) CONCEPT

5.2.1 Centre & Surround Operation in Discriminant Saliency

Discriminant Saliency (DIS) studies differences between two pixel distributions sampled from assigned windows called centre and surround windows. It intuitively fits into the concept of centre and surround operators, and the idea would work well if linear difference such as image contrast needs figuring out. However, deployment of mutual information, a high-order comparison between two distributions would require enormous amount of sampled pixels which correspondent image patches certainly can not afford. Therefore, centre-surround operators have to be considered in the informatics and statistics' point of view.

Modelling 2-D data such as images is a complicated task due to infinite number of possible joint distributions. However, in the visual attention study, natural images are the only image class which are analysed. Then, it simplifies the task significantly due to numerous works on statistics of natural images in the literature [142–147]. From several studies about statistical aspects of natural images, we notice importance of explicit and implicit manifolds by Shi and Zhu [148]. They have argued image patches as fundamental elements for objects modelling and recognition in natural images. Moreover, these patches play a key role in forming the whole ensemble of natural image patches which are in turn classified into two types of subspaces: explicit and implicit manifolds. The explicit manifold mainly contains simple and regular image primitives such as edges, bards, corners, and junctions while implicit manifold is dominated with complex and stochastic image patches such as textures and clutters. Patch scaling realizes a connection between two types of manifolds; in other words, a patch can be classified as either edge or texture according to its window size. Furthermore, image/patch scales not only change nature of image patches from implicit to explicit manifolds but as well affect their entropy since two subspaces live in separate regimes of low and high entropy . Studying this transition over scales shows the peak of manifold complexity in the middle entropy regime [148]. In other words, there exists an intermediate scale which collects the most information of patches if information is measured by entropy value.

The transition from implicit to explicit manifolds or vice verse over multiple scales in natural images have strong correlation with centre-surround operators of visual attention theory. Centre or surround windows considerably correspond to smaller and bigger patches at the same location; in other words, they are patches at two consecutive scales. When two or more pairs of centre and surround patches are employed, multi-

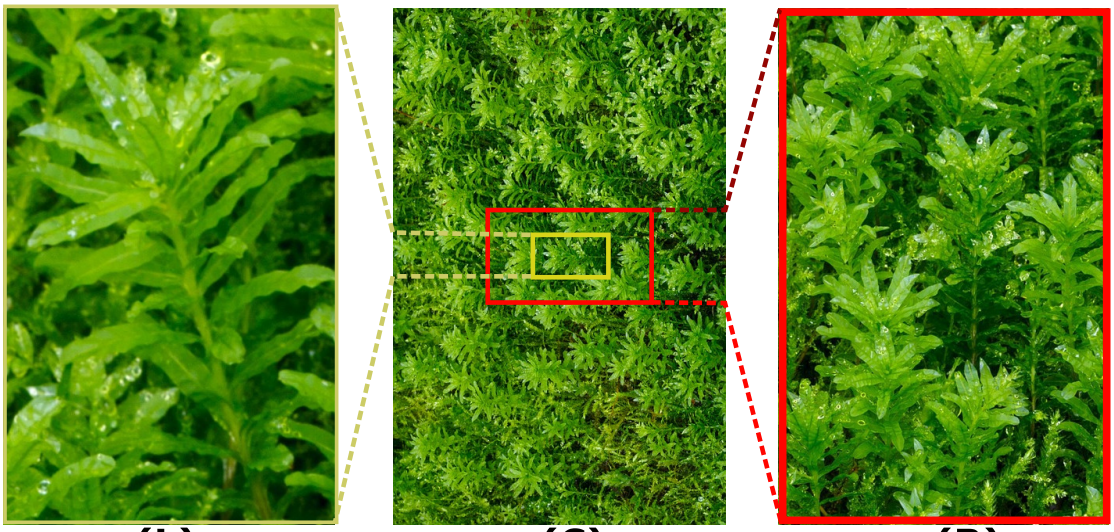


Figure 5.1: (L) : central contex (explicit manifold). (C): original image. (R): surrounding context (implicit manifold)

scale centre-surround operators are formed with blocks at different scales. As image scaling plays a key role in classification of patches into either implicit or explicit manifold, centre and surround patches have different likelihood towards the manifolds. Therefore, DIS can be formulated as binary classification between two different types of manifolds with statistical constrains instead of two distinguishing patches with geometrical constrains. It certainly benefits estimation of mutual information because DIS with statistical constrains can be learnt from patches across images, the global context while the original DIS with geometrical constraint restrains the accuracy of information estimation due to its small number of pixels in local contexts.

Deployment of DIS approach in deciding classes of image patches according to their statistical characteristics needs mathematical models of information estimation for realization. Shi and Zhu [148] propose manifold pursuits along with the introduction of implicit and explicit manifolds. The pursuit utilizes iterative approaches EM algorithms to produce approximate statistical models of image patches ($\mathbf{I}$) over the global context Ω. In their approaches, training samples are image pixels, spatial features, which is again undesirable for multi-scale extension. Alternatively, statistical model can be built to describe characteristics of wavelet features with Gaussian Mixture Models (GMM), a good statistical approximation of wavelet coefficients distributions from natural images as well as their usage of large-variance and small-variance Gaussian distributions. Again, these large/small variance or implicit/explicit distributions are just two aspects of one phenomenon assumed that Gaussian distributions are suitable for modelling features of natural images. For instances, large variance Gaussian dis-

tribution means large uncertainty or high-entropy regime where implicit manifold live on; meanwhile, small variance Gaussian distribution has small uncertainty or low entropy regime where explicit manifolds exists. Therefore, statistical constrains for DIS in the global context can utilize large/small-variance Gaussian distributions instead of implicit/explicit manifolds concepts due to its convenience in mathematical models. In this chapter , GMM concepts are seen through for modelling wavelet coefficients, spatial-scale features.

5.2.2 Discriminant Saliency Formulation

In decision theory, saliency is regarded as power for distinguishing salient and non-salient classes. Such theory allows integrate the classical centre-surround hypothesis into a statistical architecture. Based on the decision theory, the discriminant detector can work with variety of stimulus modalities, including intensity, colour, orientation and motion. Moreover, various psychophysical properties for both static and motion stimuli are shown to be accurately satisfied quantitatively by DIS saliency maps. [17]

Perceptual systems evolve for optimally producing decisions about states of surrounding environments with minimum probability of error in a decision-theoretic sense as well as minimum computational effort. In order to achieve these goals, artificial systems needs a mathematical framework for optimization. Mathematically, the problem can be defined as (1) a binary classification of interest stimuli and null hypothesis (salient against non-salient features) and (2) measurement of discriminant power from extracted visual features as saliency at each location in the visual field. The discriminant power is estimated in classification process with respect to two classes of stimuli: stimuli of interest and null stimuli of all uninterested features. Each location of visual field can be classified whether it includes stimuli of interest optimally with lowest expected probability of error. From pure computational standpoint, the binary classification for discriminant features are widely studied and well-defined as tractable problem in the literature. Moreover, the discriminant saliency concept and the decision theory appear in both top-down and bottom-up problems with different specifications of stimuli of interest [108].

The early stages of biological vision are dominated by the ubiquity of "centre-surround" operator; then, bottom-up saliency is commonly defined as how certain the stimuli at each location of central visual field can be determined against other stimuli in its surround. In other words, "centre-surround" hypothesis is a natural binary classification problem which can be solved by well-established decision theory. In this problem, classes can be defined as follows.

- Interest hypothesis: observations within a central neighborhood W_l^1 of visual fields location l.
- Null hypothesis: observations within a surrounding window W_l^0 of the above central region.

At each location, likelihood of either hypothesis depends on visual stimulus, a predefined set of features X. The saliency at location l should be measured as discriminating power of features X in W_l^1 against X in W_l^0. In other words, discriminant saliency value is proportional to distance between feature distributions of centre and surrounding classes.

Within DIS, feature responses within the windows are randomly drawn from the predefined sets of features X. Since there are many possible combinations and orders of how such responses are assembled, the observations of features can be considered as a random process of dimension d. $X(l) = (X_1(l), \ldots, X_d(l))$. Each observation is drawn conditionally on the states of hidden variable $Y(l)$, which is either centre or surround state. Feature vectors $\mathbf{x}(\mathbf{j})$ such that $j \in W_l^c, c \in \{0, 1\}\}$ are drawn from classes c according to conditional densities $P_{X(l)|Y(l)}(x|c)$ where $Y(l) = 0$ for surround or $Y(l) = 1$ for centre. The saliency S at location l, $S(l)$, is equal to the discriminant power of X for classifying observed feature vectors, which can be quantified by the mutual information between features, X, and class labels, Y.

$$S(l) = I_l(X; Y) = \sum_c \int p_{X,Y}(x, c) log \frac{p_{X,Y}(x, c)}{p_X(x) p_y(c)} dx \tag{5.2.1}$$

Though binary classification and decision theory makes discriminant saliency computationally feasible, it is only for low-dimensional data. Computer vision and visual attention need to deal with high-dimension input images especially when it involves statistics and information theory. As mentioned previously, observations of feature responses, $X(l)$, are considered as a random process in d-dimensional space. Mutual information estimation of high-dimension data encounters serious obstacles due to the curse of dimensionality as well as computational efficiency. As these problems persist, saliency algorithms would never be biologically plausible and computationally feasible. Therefore, discriminant saliency algorithms have to be approximated by taking into account statistical characteristics of natural images as well as mathematical simplification. Dashan Gao and Nuno Vasconcelos [17] have proposed a feasible algorithm for mutual estimation, mathematically formulated as follows.

$$
\begin{aligned}
I_l(X;Y) =& H(Y) - H(Y|X) && (5.2.2)\\
=& E_X(H(Y) + E_{Y|X}[\log(P_{Y|X}(c|x))]] && (5.2.3)\\
=& E_X\left[H(Y) + \sum_{c=0}^{1} P_{Y|X}(c|x)\log(P_{Y|X}(c|x))\right] && (5.2.4)\\
=& \frac{1}{|W_l|}\sum_{j\in W_l}\left[H(Y) + \sum_{c=0}^{1} P_{Y|X}(c|x_j)\log\left(P_{Y|X}(c|x_j)\right)\right] && (5.2.5)
\end{aligned}
$$

where

- $H(Y) = -\sum_{c=0}^{1} P_Y(c)\log(P_Y(c))$ is the entropy of classes Y
- $H(Y|X) = -E_{Y|X}\left[\log(P_{Y|X}(c|x))\right]$ is the conditional entropy of Y given X

Given a location l, there are corresponding centres W_l^1 and surround W_l^0 windows along with a set of associated feature responses $x(j), j \in W_l = W_l^0 \cup W_l^1$. The mutual information can be estimated by replacing expectations with means of all samples inside the joint windows W_l. A conditional entropy $H(Y|X)$ can be computed by analytically deriving MAP of $P(Y|X)$ given deployment of Generalized Gaussian Distribution (GGD) for transformed features in a binary classification problem. Let's name Gao's proposal for computation of discriminant saliency as DIS; more details about DIS can be found in their publications [108][17].

While DIS successfully defines discriminant saliency in decision theoretic sense, its implementation has certain limits. Feature responses are randomly sampled in a single fixed-size window; therefore, it is obviously biased toward objects with distinctive features fitted in that window size. As previous findings have confirmed involvement of multi-scale factors in visual attention, DIS needs extending from a fixed-scale process to a multi-scale process. In theory, DIS can be carried out with different size of windows, and this approach certainly produces image responses and saliency values at multiple scales. However, such approach is not recommended for both computational and biologically mechanisms since it causes high redundancy in saliency values across multiple scales. In order to solve the multi-scale problem systematically, DIS should be integrated with multiple scale processing techniques such as wavelet transforms.

5.3 MULTISCALE DISCRIMINANT SALIENCY (MDIS) MODELLING

Expansion from a fixed window-size to multi-scale processing is commonly desired in development of computer vision algorithms [26] [22]. In long literature of computer vision research fields, there are many multi-scale processing models [22] [149] [150] [151] which can be used as references for developing a Multi-scale Discriminant Saliency (MDIS). Any selected model has to adapt binary classification in multi-scale levels. In other words, it requires efficient classification of image data into a class at a particular scale with prior knowledge from other scales. With respect to these requirements, a multi-scale image segmentation framework should be a good starting point for MDIS since DIS can be considered as simplified binary image segmentation with only two classes. However, the binary classification is only an intermediate step to measure discriminant power of centre-surrounding features, and accuracy of segmentation results does not really matter in this case.

Typical algorithms employ a rigid classification window in a vague hope that all pixels in region of interest (ROI) belong to the same class. Obviously, DIS has similar problems of choosing suitable window sizes as well. Clearly, the size is crucial to balance between its reliability and accuracy. A large window usually provides rich statistical information and enhance reliability of the algorithm. However, it also risks including heterogeneous elements in the window and eventually loses segmentation accuracy. Therefore, appropriate window sizes are equivalently vital in avoidance of local maxima in calculation discriminant power. If window sizes are too large or too small, MDIS risks losing useful discriminative features or being too susceptible to noise. In brief, sampling rates, and consequently number of data in a window, directly affect on performance of binary classification / segmentation and eventually computation of discriminating powers.

5.3.1 Multi-scale Classification Windows

Multi-scale segmentation [150] employs multi-scale classification windows on images, then combines responses across scales. MDIS can adapt similar approach to classify image features into either centre or surrounding class for multiple scales and computer saliency values across multiple scales. In image processing field, dyadic scales (or blocks) are often implemented in MDIS because of its compactness and efficiency [26]. Let's assume an initial square image s with 2^Jx2^J of $n := 2^{2J}$ pixels, the dyadic square structures can be generated by recursively dividing x into four square sub-

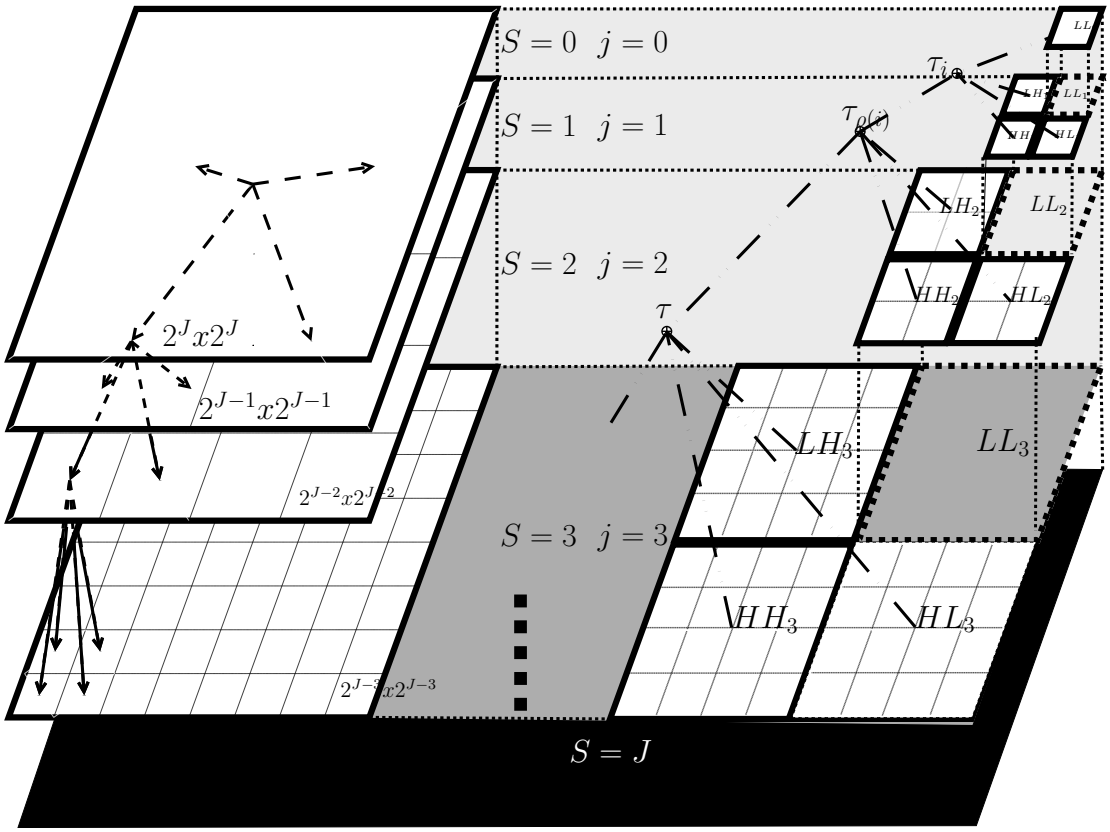


Figure 5.2: Quad-Tree and Dyadic Wavelet Structures

images equally , the left part of Figure 5.2 . As a result, it becomes the popular quad-tree structure which is also commonly employed in computer vision and image processing problems. In this tree structure, each node is related to a direct above parent node while it plays a role of parental nodes itself for four direct below nodes , the right part of Figure 5.2 . Each quad-tree node is equivalent with a dyadic square, and is denoted as a tree-node in scale j by d_i^j whereof i is a spatial index of a dyadic square node. Given a random field image X, the dyadic squares are also random fields, formulated as D_i^j mathematically. In following sections, D_i (dropping scale factor j) as general randomly-generated dyadic square regardless of scales are used. With these dyadic squares as classification windows, each node d_i can be classified as either centre or surround by estimating its maximum a posterior probability (MAP). Then, mutual information between features and corresponding labels is an average of MAP across all classes [109]. Mutual information is similar to the core concept of discriminant power for central features against surrounding ones at each location [108]. Moreover, deployment of quad-tree structures makes the estimation of mutual information possible across multiple scales feasible by learning from multiple PDFs through Hidden Markov Tree (HMT), a wavelet-based statistical model. The HMT framework's formulation, training and testing are described in following Sub-Sections 5.3.2 and 5.3.3.

5.3.2 Multi-scale Statistical Model

Hidden Markov Tree (HMT) [139] [149] aims to capture joint PDF of wavelet coefficients across multiple scales $f(w|\Theta)$, where $\mathbf{w}$ is a vector of coefficients across multiple scales; Θ are correspondent HMT model parameters. How these parameters can be identified .i.e through training processes or prior knowledges leads to multiple derivatives of HMT. These derivatives are studied in this section for their advantages and limitations.

Modelling marginal distributions of wavelet coefficients w_i benefit their sparseness in representing natural images since a minority of coefficients are large while a majority of them are small. That distribution is efficiently captured by Gaussian Mixture Model (GMM) with wavelet coefficients w_i and observed hidden state variables or class labels $S_i \in \{S, L\}$. Assumed that Gaussian Probability Distribution Function (PDF) is mathematically described as follows

$$g(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi}\sigma} exp - \frac{(x-\mu)^2}{2 * \sigma^2} \tag{5.3.1}$$

where x is randomly distributed variable, μ is mean, and σ is variance, lets denote the Gaussian PDF of small-variance state $S_i = S$ as follows

$$f(w_i|S_i = S) = g(w_i; 0, \sigma_{S_i}^2) \tag{5.3.2}$$

While Gaussian PDF of large-variance state $S_i = L$ is formulated as follows

$$f(w_i|S_i = L) = g(w_i; 0, \sigma_{L_i}^2) \tag{5.3.3}$$

where $\sigma_L^2 \geq \sigma_S^2$. By those mixture models, the marginal PDF $f(w_i)$ with respect to GMM can be written as a convex combination of the conditional densities.

$$f(w_i) = p_i^S g(w_i; 0, \sigma_{S_i}^2) + p_i^L g(w_i; 0, \sigma_{L_i}^2) \tag{5.3.4}$$

where $p_i^S + p_i^L = 1$ since $p_{S_i} = [p_i^S p_i^L]$ are mass probability of states $\{S, L\}$. In statistical interpretation, p_{S_i} is how likely wavelet coefficients w_i are small S or large L.

HMT is designed to capture inter-scale statistical dependencies between hidden state variables of wavelet coefficients and those of its four children. These hidden state variable are called *nodes* of HMT probabilistic trees which connect wavelet coefficients across all levels. The probabilistic tree of hidden state variables or dependency graph have similar quad-tree topology as wavelet decomposition, as illustrated in the right

side of the Figure 5.2. The only difference is that each node represents a hidden variable for a wavelet coefficient instead of the coefficient itself. Therefore, HMT can model the each wavelet coefficient by a two-states $\{S, L\}$ of Gaussian Mixture Model (GMM), as mentioned in Equation 5.3.4. Moreover, HMT can also model joint PDF of a pair of parent and child coefficients by mathematically formulating their state-to-state links as follows.

$$A_i = \begin{bmatrix} p_i^{S\to S} & p_i^{S\to L} \\ p_i^{L\to S} & p_i^{L\to L} \end{bmatrix} \tag{5.3.5}$$

where $p^{S\to S}$ or $p^{L\to L}$ describes a probability for which hidden states of both parent and child wavelet coefficients are them same, both small S or large L. These probabilities $p^{S\to S}$ and $p^{L\to L}$ are called persistency probabilities. Meanwhile, $p^{S\to L}$ or $p^{L\to S}$ describes a probability for which a hidden state is changed from small S to large L or vice versa. These probabilities $p^{L\to S}$ and $p^{L\to S}$ are called novelty probabilities . Persistency probabilities are lying on the main diagonal axis of the above array, as shown in Equation 5.3.5 , $p_i^{S\to S}$,$p_i^{L\to L}$, and they represent how likely states are preserved in parent and child links. On the sub diagonal axis are novelty of probabilities, which causes different states between parent and child nodes. In addition, the relationship between the persistency probability and novelty probability are described as $p_i^{S\to S} + p_i^{S\to L} = 1$ and $p_i^{L\to S} + p_i^{L\to L} = 1$ since these probabilities forms two possible combinations of parent and child state-to-state links $\{S \to S, S \to L\}$ or $\{L \to S, L \to L\}$.

In brief, Hidden Markov Tree can model statistical relationships of wavelet coefficients across multiple dyadic scales through three below probabilities.

1. Probability of each wavelet coefficient whose hidden state variable is either small S or large L with respect to GMM model.

2. Probability of each parent and child wavelet coefficient pair whose state-to-state links are either $\{S \to S, S \to L\}$ or $\{L \to S, L \to L\}$.

3. Initial probability of wavelet coefficients at the coarsest level p.

Therefore, the HMT model can be identified by following parameters.

1. $\sigma_{S;i}^2$ and $\sigma_{L;i}^2$ of GMM, as shown in Equations 5.3.2 and 5.3.3

2. A_i, transition of states matrix, as shown in Equation 5.3.5

3. p, a probability mass function at the coarsest level.

By grouping these parameters, a HMT can be defined by a parametric model with parameters Θ

$$\Theta = \{p, A_{b,j,k}; \sigma_{S_i;k,j,b}\} \tag{5.3.6}$$

where

$$\begin{aligned} j &= 1, \ldots, J \\ S_i &= \{L, S\} \\ k &\in \mathbb{Z}^2 \\ \forall b &\in \mathcal{B} \end{aligned}$$

where j is level index, k is spatial index of wavelet coefficients while b represents index of three wavelet sub-bands (High-Low, Low-High, and High-High) $\mathcal{B} = \{HL, LH, HH\}$ in Discrete Wavelet Transform (DWT) assumed that wavelet transform is carried out for J levels. The parametric model Θ_c can be learned from the joint PDF of wavelet coefficients in each of three sub-bands $\mathcal{B}$ [139]. A generality of a parametric model allows a different set of parameters for each node. However, it overcomplicates the model with too many parameters since n nodes are required to be fitted on $4n$ parameter models [139]. It is not computationally feasible task due to curse-of-dimensionality. Therefore, different from a general parametric model, HMT is assumed to use same parameters for every node at the same scale of the wavelet transform. This simplification can be expressed through changes in notations of HMT parameters, as shown Equation 5.3.6.

$$\sigma^2_{S;b,j,k} = \sigma^2_{S;j} \tag{5.3.7}$$

$$\sigma^2_{L;b,j,k} = \sigma^2_{L;j} \tag{5.3.8}$$

$$A_{b,j,k} = A_j \tag{5.3.9}$$

Variances σ and transformation matrices A become independent of spatial k and oriental b indexes. The assumption is called *tying within scale* or shortly *tying* [139] and it makes HMT model less complex and computationally feasible but less general. In accordance with this *tying* constrain, HMT model is mathematically formulated as follows.

$$\Theta_{THMT} = \{p, A_j; \sigma_{S_i;j}\} \tag{5.3.10}$$

where

$$\begin{aligned} j &= 1, \ldots, J \\ S_i &= \{L, S\} \end{aligned}$$

Lets name the Θ_{THMT} in equation 5.3.10 as trained HMT (THMT) as these parameters need to be trained and learn from input data. While the *within scale* constraint has significantly simplified HMT parameters, which in turns reduces complexity of a training process. In fact, parameters of THMT can be trained with data from a single input image. However, it is possible to eliminate the training process for HMT parameters if common characteristics of an image category is known in advance and used as prior information for a HMT model. HMT can be trained at once for parameters with a few images in the same category and these meta-parameters are assumed to be HMT parameters of all other images from this category as well. In other words, it yields a general HMT model for that class of images with each member in the class being treated statistically equivalent [149]. Fortunately, Romberg [149] have studied similar models for the class of natural images and published this universal parameters of the HMT model Θ_{UHMT} for natural images [149]. This approach of modelling HMT is called Universal Hidden Markov Tree (UHMT) in this chapter . As MDIS aims to model interest rate of data from natural images; therefore, it is reasonable to utilize this HMT model with the universal parameters [149] for MDIS. Though this UHMT model does not closely reflect statistical characteristics of all different images, its assumption totally eliminates training processes and make real-time HMT possible. Experiments with UHMT mode are mentioned in the Section 5.4 where both accuracy and efficiency of UHMT are evaluated.

While *'tying'* THMT and UHMT approach simplify training stages of HMT, these models are less general and unable to model images with complex textures. *tying* THMT assumes the same parameters for wavelet sub-bands disregards any statistical relations of coefficients across different orientations. However, experiments by Simoncelli and Portilla [152] demonstrate importance of cross-correlation between wavelet sub-bands or coefficients from different orientations at the same scale for modelling texture image. Since the above THMT and UHMT treat coefficients of sub-bands independently, they obviously ignore the cross-orientation correlation. To enhance the capacity of THMT, Do and Vetterli [146] propose grouping of coefficients at the same location and scale into a vector, then carry out HMT modelling in a multidimensional manner. In this chapter, such approach is called Vectorized HMT (VHMT). As textural features are important in a class of natural images as well; therefore, capturing this dependency with VHMT would improve accuracy of MDIS. Let's denote a vector after the grouping of wavelet coefficients at location k, scale j in three different orientations vertical (V), horizontal (H), and diagonal (D) as follows.

$$\mathbf{w}_{j,k} = (w^H_{j,k} w^V_{j,k} w^D_{j,k})^T \tag{5.3.11}$$

As multivariate wavelet coefficients $\mathbf{w}_{j,k}$ are concerned, their distribution needs to be formulating as zero-mean multivariate Gaussian density with covariance matrix C instead of a variance σ as follows

$$g(\mathbf{w}; 0, C) = \frac{1}{\sqrt{(2\pi)^n |det(C)|}} \exp(-\mathbf{w}^T C^{-1} \mathbf{w}) \tag{5.3.12}$$

where n is the number of dimensions, or orientations $n = 3$ in this case; C is covariance matrix of multivariate Gaussian distribution. Except from a multivariate probability density in Equation 5.3.12, VHMT is similar to THMT. For example, its marginal distribution is formulated as Gaussian Mixture Models, as shown in Equation 5.3.13 and VHMT describes links between states of parent and child nodes with the same transition matrix, A, as shown in Equation 5.3.5.

$$f(\mathbf{w}) = p^L g(\mathbf{w}; C^L) + p_j^S g(\mathbf{w}; C^S) \tag{5.3.13}$$

Where p^L and p^S are mass probability of states $\{S, L\}$ and $p^L + p^S = 1$. Moreover, its statistical inter-scale dependency is modelled through the parent-child relationship with a quad-tree structure linking a parent with its four children at the next level in the same location, as illustrated in the Figure 5.2. A small difference here is that only one tree is utilized instead of three trees since wavelet coefficients have been grouped and modelled simultaneously. Hence, an image can be modelled by VHMT, Θ_{VHMT} with a set of parameters.

$$\begin{aligned} \Theta_{VHMT} =& \{p, A_j; C_{S_i;j}\} \\ &\text{where} \\ j =& 1, \ldots, J \\ S_i =& \{L, S\} \end{aligned} \tag{5.3.14}$$

Where p is an initial probability at the coarsest level of VHMT, A_j, as defined in the Equation 5.3.5, is state-to-state transition matrix at the level j and $C_{S_i;j}$ are covariance factor of a multivariate Gaussian distribution. As only one quad-tree is built for modelling a vector of wavelet coefficients, the hidden states are as well simplified by 'tying within scale constrain'. It means the same hidden state is assigned regardless of orientations; in other words, VHMT is orientation-invariant. VHMT captures dependencies across orientations via a covariance matrix C of a multivariate Gaussian density. Diagonal elements of the C matrix are variances of each orientation meanwhile non-diagonal elements represent covariance of wavelet coefficients across sub-bands. It justifies the

VHMT model for textural features because their wavelet coefficients have high possibility of being significant at all orientations in edge regions whereas they are likely small at any directions in smooth regions.

5.3.3 Multi-scale Statistical Learning

The complete joint PDF from image pixels is typically overcomplicated and difficult to model due to their high-dimensionality. Unavailability of simple distribution model in practice motivates statistical modelling and learning from image data in wavelet-domain. In this transform domain, image representation are often more sparse, less complex; therefore, simple distribution models are applicable. Hence, joint pixel PDF could be well approximated as marginal PDF of wavelet coefficients. Moreover, wavelet analysis well-characterizes semantic singularity or edges of natural images, it provides a suitable transform-domain for modelling statistical property of these singularity-rich images.

Natural images are full of edges, ridges and other highly structural features as well as textures; therefore, DWT wavelet transform, which are able to capture the singularities across multiple scales and different directions, behaves as a multi-scale edge detector. Though only discrete wavelet transform (DWT) is considered in this chapter; this statistical learning concept can be adapted into other types of wavelet transforms as well. Henceforth, whenever a wavelet transform is mentioned in this chapter, it refers to DWT unless otherwise stated. As multi-scale edge detectors, responses of wavelet transform overlying a singularity are large coefficients while DWT analysis of a smooth region yields small coefficients. Simple hard-threshold of wavelet coefficients leads to binary differentiation of singularity against non-singularity features. Moreover, statistical model of images can be significantly simplified under "restructured" multi-scale singularity representation or the quad-tree structure of parent and child wavelet-coefficients as illustrated in the Figure 5.2. Quad-tree structure of dyadic squares in pixel domain, as illustrated in the left-hand side of the Figure 5.2, can be mirrored into wavelet decomposition, as shown in the right part of the Figure 5.2 since four wavelet coefficients at a given scale nest inside the coefficient at the next coarser level. The combination of singularity detector and multi-scale quad-tree structure implies a singularity property at each spatial location persists through scales along the branches of the quad-tree in the transform domain. The singularity characterization along scales makes the wavelet domain well-suited for modelling natural images. In this chapter, statistical relations of wavelet coefficients are modelled by HMT model Θ [139] [149], as described in the previous Sub-Section 5.3.2. Parameters of the HMT

model can be learnt by Estimation - Maximization (EM) algorithms.

Algorithm 5 EM Algorithms for HMT Training

Require: w_{tr} : wavelet coefficients for training
  Initialize a set of parameters Θ_0, hidden state variable S
  for step $l = 0 \ldots \infty$ **do**
    E-step: compute the expected value $E_S(\log f(w_{tr}, S|\Theta)|w, \Theta_l)$
    M-step: update Θ_{l+1} such that E_S is maximized.
      $\Theta_{l+1} = \arg\max_{\Theta}(E_S(\log f(w_{tr}, S|\Theta)|w, \Theta_l))$
  end for
Ensure: Θ_{HMT} .s.t $\Theta_{HMT} = \arg\max_{\Theta}(\log f(w_{tr}|\Theta))$

The parameters of a HMT model Θ can be learned through an iterative expectation and maximization (EM) algorithm with cost $\mathcal{O}(n)$ per iteration where n is the number of nodes [139]. While EM training is simple for implementation and gives a reasonable estimate of parameters, it has following draw-backs .i.e convergence to local maximum of log-likelihood $\log f(w|\Theta)$ and high computational expense. Moreover, though training iteration does not last for infinity, there is no upper-bound for number of iterations before convergence. After the parameters Θ_{HMT} are estimated by the EM iteration, statistical characteristics of a DWT coefficient w_{te} can be computed by Equation as their log-likelihood given the HMT model.

$$p(w_{te}|\Theta_{HMT}) = \log f(w_{te}|\Theta_{HMT}) \tag{5.3.15}$$

Moreover, computation of this likelihood $\log f(\mathbf{w}|\Theta_{HMT})$ for all wavelet coefficients across multiple levels and sub-bands requires only a simple "up-sweep" operation through the HMT tree from leaves to root [139].

Conveniently, HMT models use similar structures as wavelet decomposition and dyadic squares do. Therefore, statistical behaviours of each square block $\mathbf{d_i}$ can be approximately computed by a HMT branch rooted at a node i, as shown in the Figure 5.2. Availability of log-likelihood for a square block inspires a simple formulation of multi-scale centre-surround operations. Supposed centre and surround classes are denoted as $c \in \{1, 0\}$ and specified or trained HMT trees for each class with parameters Θ_c is achieved, the likelihood calculation is computed by the Equation 5.3.15 for each node of the HMT quad-tree with "up-sweep" operation given the wavelet coefficients w. For each node of the tree or each dyadic block $\mathbf{d_i}$, HMT yields likelihood $f(\mathbf{d_i}|\Theta_c)$ given $c \in \{1, 0\}$ for each dyadic block $\mathbf{d_i}$. Therefore, the block $\mathbf{d_i}$ can be classified as centre or

surround label, the most suitable class c, according to which class has larger likelihood.

$$\hat{c}_i^{ML} := \arg\max_{c \in \{1,0\}} f(\mathbf{d_i}|\Theta_c) \tag{5.3.16}$$

The most likely label $\hat{c}_i^{ML}$ for each dyadic sub-square $\mathbf{d_i}$ can be found by a simple comparison between estimation of available classes. Moreover, labels of all pixels can be found by this approach with an "up-sweep" operation from leave-node to root node along the Hidden Markov Tree.

5.3.4 Multiscale Likelihood Computation

Given $\mathbf{w_i}$ is a vector of wavelet coefficients which is analysed by multiple levels from a block $\mathbf{d_i}$ and HMT parameters Θ are learnt with the coefficients $\mathbf{w_i}$, the likelihood of this block can be computed as follows

$$f(\mathcal{T}|\Theta) = \sum_{m=S,L} \beta_i(c) p(S_i = m|\Theta) \tag{5.3.17}$$

Where $\mathcal{T}_i$ denotes a set of all wavelet coefficients of a HMT tree $\mathbf{w_i}$, a conditional likelihood $\beta_i(m) := f(\mathcal{T}_i|S_i = m, \Theta)$ is computed with Equation 5.3.15 by sweeping up coefficients from leave-node to root node of the HMT tree [139], and a state probability $p(S_i = m|\Theta)$ can be predefined in UHMT or trained for in (T/V)HMT [149]. In DWT analysis, a block $\mathbf{d_i}$ is represented by not only one but three sub-trees, $\{\mathcal{T}_i^{LH}, \mathcal{T}_i^{HL}, \mathcal{T}_i^{HH}\}$ corresponding to three sub-bands $\mathcal{B} = \{LH, HL, HH\}$. Likelihood of each sub-band can be independently calculated by the Equation 5.3.17 with coefficients $\mathbf{w}$ from their corresponding tree $\mathcal{T}$. This independence is appropriate computation step assumed that correlation between feature channels would not affect the likelihood. In turn, it does not alter decision of centre or surround label for the block; in other words, the centre-surround discriminant power [17]. The assumption is reasonable since DWT is known as de-correlation tools and decorrelated signals are linearly independent from each other. Hence, the likelihood of a dyadic square is formulated as product of three independent likelihoods of wavelet sub-bands at each scale.

$$f(\mathbf{d_i}|\Theta) = f(\mathcal{T}_i^{LH}|\Theta^{LH}) f(\mathcal{T}_i^{HL}|\Theta^{HL}) f(\mathcal{T}_i^{HH}|\Theta^{HH}) \tag{5.3.18}$$

where Θ is HMT parameters for the whole block $\mathbf{d_i}$; Θ_{LH}, Θ_{HL} and Θ_{HH} are HMT parameters of corresponding sub-trees The assumption of independent sub-bands is only necessary in the (U/T) HMT model while VHMT has grouped wavelet coefficients into multivariate variables. Therefore, it only needs one quad-tree representation of

multivariate coefficients, $\mathcal{T}$, and likelihood of dyadic squares under each tree node is formulated as follows.

$$f(\mathbf{d_i}|\Theta) = f(\mathcal{T}_i|\Theta) \tag{5.3.19}$$

With equation 5.3.18, it is able to identify likelihood of a block $\mathbf{d_i}$ as small as 2x2 since the smallest size of a discrete mother wavelet is 2x2. Moreover, only wavelet-coefficients are involved in HMT models; while sub-bands LL, a low-passed approximation of original images, are not utilised. Without sub-band LL, pixel brightness and lightning conditions of scenes are ignored in modelling. With sub-band LL, the likelihood can be computed for even 1x1 block or a pixel.

The formulation of likelihood in Equations 5.3.18 and 5.3.19 only gives "raw" decision of centre or surround class according to available data in the block. This classification decisions is lack of inheritance across dyadic scales due to isolation the likelihood estimation of wavelet coefficients at each scale from other scales or likelihood estimation of a block from neighbouring blocks. Therefore, a better classification decision (centre / surround) can be achieved by integrating prior knowledge of other scales or at least the direct coarser scale. Details are presented in the next Sub-Section 5.3.5.

5.3.5 Multi-scale Maximum a Posterior

In the previous section, only "raw" binary classification between two states has been made under the wavelet HMT model [139]. Given a prior knowledge from other scales, a better centre-or-surround classification can be made for DIS and MDIS, as shown the Equation 5.2.5 by maximizing a posterior probability (MAP) $f(c_i^j|\mathbf{d_i^j})$ whereof c_i^j and $\mathbf{d_i^j}$ are respectively class labels and features of a block at a dyadic scale j and location i.

In order to estimate the MAP , Bayesian approach for capturing dependencies between dyadic blocks at different scales need to be employed. Though many approximation techniques [153],[154],[155] are derived for a practical computation of MAP, the Hidden Markov Tree (HMT) by Choi [150] is proven to be a feasible solution. Choi [150] introduces hidden label tree modelling instead of joint probability estimation in high-dimensional data of dyadic squares. Due to strong correlation of a square under inspection with its parents and their neighbours, class labels for these adjacent squares would affect the class decision at the considered block. For example, if a parent block belongs to a certain class and so do their neighbours, the child block most likely belongs to the same class as well. Therefore, the parent-child relation should be modelled by a general probabilistic graph [139]. However, the complexity exponentially increases with

number of neighbouring nodes. Choi [150] proposes an alternative simpler solution, based on context-based Bayesian approach. For the sake of simplicity, causal contexts are only defined by states of the direct parent node and its 8 intermediate neighbours. Let's denote the context for $\mathbf{d_i}$ as $\mathbf{v_i} \equiv [v_{i,0}, v_{i,1}, \ldots, v_{i,8}]$ where $v_{i,0}$ refers to labels of the direct parent node and their neighbours. The triple $\mathbf{v_i} \rightarrow c_i \rightarrow \mathbf{d_i}$ forms a Markov-1 chain, relating prior context $\mathbf{v_i}$ and node features $\mathbf{d_i}$ to classification decisions c_i. Moreover, class labels of prior contexts $\mathbf{v_i}$ are chosen as discrete values as it simplifies the modelling considerably. Given that prior context, independence can be assumed for label classification at each node; therefore, it is allowed to write.

$$p(\mathbf{c^j}|\mathbf{v^j}) = \prod_i p(c_i^j|\mathbf{v_i^j}) \tag{5.3.20}$$

The property of Markov-1 chain assumes that $\mathbf{d_i}$ is independent from $\mathbf{v_i}$ given c_i; therefore, the posterior probability of classifying $\mathbf{c^j}$ given $\mathbf{d^j}, \mathbf{v^j}$ is written as follows.

$$p(\mathbf{c^j}|\mathbf{d^j}, \mathbf{v^j}) = \frac{f(\mathbf{d^j}|\mathbf{c^j})p(\mathbf{c^j}|\mathbf{v^j})}{f(\mathbf{d^j}|\mathbf{v^j})} \tag{5.3.21}$$

As independence is assumed for label decisions in classification processes, it yields

$$p(\mathbf{c^j}|\mathbf{d^j}, \mathbf{v^j}) = \frac{1}{f(\mathbf{d_i^j}|\mathbf{v^j})} \prod_i f(\mathbf{d^j}|c^j)p(c^j|\mathbf{v^j}) \tag{5.3.22}$$

Hence, the marginalized context-based posterior is written as follows.

$$f(c_i^j|\mathbf{d_i^j}, \mathbf{v_i^j}) \propto f(\mathbf{d_i^j}|c_i^j)p(c_i^j|\mathbf{v_i^j}) \tag{5.3.23}$$

It greatly simplifies MAP posterior estimation which no longer needs to deal with joint prior conditions of features and contexts. It only needs to obtain two separated likelihoods of the dyadic square given the class value c_i, $f(\mathbf{d_i}^j|c_i^j)$, and the prior context provided through $\mathbf{v_i}$, $p(c_i^j|\mathbf{v_i^j})$.

While retrieving the likelihood $f(\mathbf{d_i}^j|c_i^j)$ is straightforward by up-sweeping operations, as shown in Equations 5.3.18 or 5.3.19 with given HMT model parameters at each scale, the complexity of prior context estimation greatly depends on its structures. Though more general structures may give better prior information for classification, it also greatly complicates the modelling and summarizing information conveyed by $\mathbf{v_i^j}$ as well. In other words, the constructed model is on the verge of context dilution, especially in case of insufficient training data [153],[139].

To simplify but still guarantee generalization of prior information, a simple context

structure, inspired by the hybrid tree model [155] in context-labelling trees, is employed. Instead of including labels from all neighbouring blocks, the simplified context only involves labels from the parent square $c_{\rho(i)}$ and major vote of class labels from the neighbouring squares $c_{\mathcal{N}(i)}$. As there are only two class labels $c := \{0,1\}$, the prior context $\mathbf{v_i} := \{c_{\rho(i)}, c_{\mathcal{N}(i)}\}$ can only been drawn from $N_c^2 = 4$ permuted sets of binary values $\{0,0\},\{0,1\},\{1,0\},\{1,1\}$. Despite such ad-hoc simple contextual model, it provides sufficient statistic for demonstrating effectiveness of a multi-scale decision fusion process [155]. Another advantage of the context structure simplification is not requiring enormous number of training data for probability estimation.

In brief, different from "raw" classification shown in Equation 5.3.16, decision of block labels at a scale j depends on not only features of a block $\mathbf{d_i}$ but also prior information of labels of neighbouring blocks $\mathbf{v_i}$ on a scale $j-1$. The posterior probability in Equation 5.3.24 can be maximized in multi-scale coarse-to-fine manner by fusing the likelihood $f(\mathbf{d_i}|c_i)$ given the label tree prior $p(c_i^j|\mathbf{v_i})$. The fusion step helps pass down MAP estimation through scales which enhances coherency between classifying results across consecutive scales. A decision about class label c_i^j can be mathematically defined by MAP given features of a block $\mathbf{d_i}$ and the prior context $\mathbf{v_i}$ as follows.

$$\hat{c_i}^{MAP} = \arg\max_{c \in 0,1} f(c_i^j|\mathbf{d^j}, \mathbf{v^j}) \tag{5.3.24}$$

5.3.6 Multiscale Discriminant Saliency

Core ideas of DIS and MDIS are measuring discriminant power between two classes centre and surrounds. Though it can be estimated by sample means of mutual information, the underlying mechanism is distinguishing the centre and surround classes given Generalized Gaussian Distribution (GGD) of wavelet coefficients. These distributions are usually zero-mean and well-characterized with only variance parameters. Dashan Gao [17] has estimated scale parameter (variance) of GGD (see section 2.4 [17] for more details) by the maximum a probability process.

$$\hat{\alpha}^{MAP} = \left[\frac{1}{\mathcal{K}}\left(\sum_{j=1}^{n}|x(j)|^{\beta} + \nu\right)\right]^{\frac{1}{\beta}} \tag{5.3.25}$$

The above MAP formula is then used for deciding whether a sample point or an image data point belongs to either the centre or surround class (see [17] for a detailed proof and explanation). Then, the more distinguishing MAP estimation of the centre class's variance parameter α_1 is from that of the surround class's variance parameter α_0, the

more discriminant power for classifying interest from null hypothesis is.

Beside GGD, Gaussian Mixture Model (GMM) is a popular choice for modelling wavelet distributions with variances of multiple classes as well [156],[157],[158]. In binary classification with only two classes, GMM includes two Gaussian Distribution mixtures (GD) of different variances, which are named as "large" / "small" states according to their comparison in terms of variance values. Now the only difference between GMM models and Gao's proposal [108] are whether GD or GGD should be used. Though GGD is more sophisticated with customizable distribution shape parameter β, several factors support validity of simple GD modelling given the class conditions as hidden variables. Empirical results from estimation have shown that the mixture model is simple yet effective [155].

For wavelet coefficients, hidden classes of "small"-"large" variance states or centre-surround states can be well modelled by Wavelet-based Hidden Markov Model (HMT) [139]. As statistical characteristics of image blocks are extracted in a coarse-to-fine multi-scale manner, MAP of a state c_i^j given input features from these blocks d_i^j can be inherently estimated across scales $j = 0, 1, \ldots, J$. Details about this multi-scale MAP estimation by wavelet HMT can be found in the previous section 5.3.5. Then, a combination of MAP estimation, in the Equation 5.3.24 and mutual information computation, in Equation 5.2.5, yields the MDIS mathematical formulation.

$$I_i^j(c_i^j; \mathbf{d_i^j}, v_i^j) = H(c^j) + \sum_{c=0}^{1} f(c_i^j|\mathbf{d_i^j}, v_j^i) log f(c_i^j|\mathbf{d_i^j}, v_j^i) \tag{5.3.26}$$

where $H(c^j) = -p(c^j)log(p(c^j))$ is an entropy estimation of classes across scales j, and the posterior probability can be estimated by the Equation 5.3.24. As the equation 5.3.26 yields discriminant power across multiple scales; a strategy is needed for combining them across scales into a saliency value. MDIS selects the maximum discriminant value across scales as its saliency value for a block $\mathbf{d_i}$ in a saliency map.

$$I_i = max\left(I_i^j(c_i^j; \mathbf{d_i^j}, v_i^j)\right) \tag{5.3.27}$$

5.4 DISCUSSION & RESULT

In the light of decision theory, saliency maps are considered as binary filters applied at each image location. According to a certain saliency threshold, each spot can be labelled as interesting or uninteresting. If a binary classification map is considered as saliency map which leads to visual performance of human beings, it would be a

significant undervaluation of human visual attention system. Psychological experiments show much better capacity of biologically plausible visual attention system than that of binary classification maps. Therefore, the proposed method deploys such classification between centre-surround environment as an intermediate step to develop information-based saliency map. At each location, mutual information between distributions of classes and features shows strength of discriminant power between interesting vs non-interesting classes given input dyadic sub-squares. From discrete binary values, saliency representation has continuous ranges of discriminant power. Inevitably, the generated saliency maps become more correlated to the results of human visual attention maps. To evaluate saliency representation, reliable ground truth data are utmost necessary. As the main research purpose is deepening knowledge about multi-scale discriminant saliency approaches and human visual attention relation, the ground truth data must be gotten from psychological experiments in which human subjects look at different natural scenes and their responses are collectively acquired. Moreover, the research scope only focuses on bottom-up visual saliency, the early stage of attention without interference of top-down knowledge and experiences. Human participants should be naive about aims of experiments and should not know contents of displaying scenes in advance. After these prerequisites are satisfied, human responses on each scene can be accurately collected through eye-tracking equipments. It records collection of eye-fixations for each scene, and these raw data are basic form of ground truths for evaluating efficiency of saliency methods.

In order to standardize the evaluating process, this chapter only utilize one of the most common and accessible database and evaluation tools in visual attention fields in information-based saliency studies, Niel Bruce database [4]. While proposing his InfoMax (AIM) approach, the first information-based visual saliency, he has simultaneously released its testing database as well. The reasonably small collection with 120 different colour images which are tested by 20 individuals. Each subject observes displayed images in random orders on a 12 inch CRT monitor positioned 0.75 m from their location for 4 seconds with a mask between each pair of images. Importantly, no particular instructions are given except observing the image.

Above brief description clarified validity of this database for the experiments. DIS [108] is obviously the most similar approach to the proposed MDIS, it should as well be used as a reference method. Though the original implementation DIS [108] is not available for setting up the comparison, pseudo-DIS can be simulated by MDIS approach with 1x1 window size and zero-mean distribution of pixel values. Besides DIS, the AIM method is also involved as a reference approach against which we compare the pro-

posed saliency solution MDIS in terms of performance, computational load, etc. AIM derives saliency value from information theory with slightly different computation, self-information instead of mutual-information in MDIS or DIS. Therefore, it would be considered the second best as referenced method for the later evaluation of MDIS.

Besides an appropriate database and referenced methods, proper numerical tools are also necessary for analysing simulation data. As regards to fairness and accuracy, a set of measurements LCC, NSS, and AUC recommended by Ali Borji, Sihite, Itti [159] is utilized since evaluation codes can be retrieved freely from their website [1]. Rationales behind these evaluation scores ensure reliability of the quantitative observations and conclusions.

As mentioned previously, AIM and DIS are chosen as the referenced informative saliency methods. They are selected due to the well-established reputation as well as freely accessible code and experiment database [2]. Due to multi-scale natures of the proposed MDIS, saliency maps for each dyadic-scale levels can be extracted as well as the final MDIS saliency, integrated across scales by the equation 5.3.26. The availability of such saliency maps allows evaluation of discriminant power concept for saliency in scale-by-scale manner or on the whole. HMT0 denote integrated MDIS while separated saliency maps are named as HMT1 to HMT5 in accordance with coarse-to-fine order. Through examination of MDIS in different aspects, it would be possible to observe its effectiveness in predictions of eye-fixation points and how selection of classifying window sizes might affect its performance. In addition, comparisons against AIM and DIS would contribute a general view how MDIS performs against a well-known information-based saliency method.

In the proposed methods, Hidden Markov Tree (HMT) plays a role of modelling statistical properties of images. It extracts model parameters at each scale by considering distribution of features given hidden variables (centre-surround labels in this method). Therefore, training is a necessary step before model parameters can be approximated in THMT. Normally different wavelet sub-bands are trained independently assumed no-correlation exists between them. However, HMT parameters can be acquired by grouping wavelet coefficients and training these multivariate variables in VHMT models as well. Furthermore, training steps are not always necessary if input images are all from a specific category; for example, natural images. Romberg, Choi, Baraniuk [149] have studied this Universal Hidden Markov Tree (UHMT) for the natural image class. In other words, model parameters can be fixed without any training efforts; the

[1] https://sites.google.com/site/saliencyevaluation/
[2] http://www-sop.inria.fr/members/Neil.Bruce/

approach would greatly reduce computational load for MDIS. Each configuration of the HMT model has its own advantages and disadvantages; therefore, a simulation is necessary to compare performances of (U/T/V) HMTs in terms of LCC, NSS, AUC and TIME (the computational time).

In quantitative assessment, general ideas can be drawn about how the proposed algorithms perform in average. However, such evaluation method lacks of specific details about successful and failure cases since the information has been averaged in quantitative method. In an effort of looking for pros and cons of the algorithm, qualitative evaluations are performed on saliency maps generated by MDIS in multiple scales. Furthermore, AIM,DIS saliency maps are generated and compared with MDIS maps for advantages and disadvantages of each method.

5.4.1 Quantitative Evaluation

After the above review of how simulations are built and evaluated in the previous section, following are data representation and analysis of the conducted experiments. In this chapter , six dyadic scales are deployed for any HMT training and evaluation; therefore, there are simulation modes from (U/T/V)HMT(0-6) of MDIS depending on whether training stages are deployed (T/V)HMT or universal parameters are used (UHMT). Besides U/T/V prefixes, there are suffixes (1-5) for consecutive deployments of block sizes from 32x32 down to 2x2. MDIS saliency maps could be combined according to the maximization of mutual information rule, the equation 5.3.26; therefore, there is an additional suffix 0 for the across-scale integrated saliency maps. Moreover, whenever a 1x1 window size (a pixel) is used, generated saliency maps are also considered as pseudo DIS saliency maps since MDIS with a pixel-size window is principally equivalent to DIS method in term of resolutions[108]. (U/T)HMT6 are assigned to pseudo-DIS saliency maps, as replacements for the original DIS implementation of which source code is not available . In addition to DIS, AIM is involved in simulations as the reference method. LCC, NSS, AUC and TIME, as described in Sub-Section 2.3.2, are chosen as numerical evaluation tools. Below are three tables of simulation results. Table 5.2 shows experimental data of all UHMT modes while Table 5.3 summarizes data of all THMT modes, and table 5.4 list evaluation results of all VHMT modes.

Figures 5.3a,5.3c,5.3e display ROC curves of (U/T/V)HMT with AIM as reference curves, in which solid green lines represents referenced methods such as AIM or DIS (HMT6); while, blue, red, and orange colours represent the proposed HMT methods. AIM and DIS perform better than HMT with large sampling windows .i.e 32x32 blocks in HMT1, 16x16 blocks HMT2, 8x8 blocks in HMT3; meanwhile, HMTs surpass AIM and DIS in

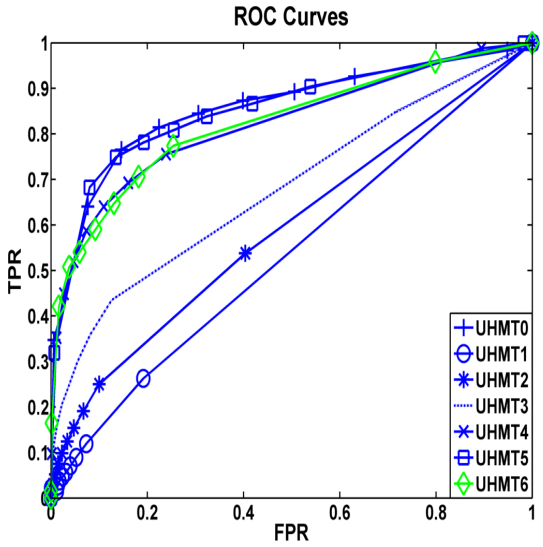
detection of fixation points with smaller block sizes .i.e 4x4 blocks in HMT4 and 2x2 blocks in HMT5. It can be summarized as follows.

$$HMT0 \geq HMT5 > DIS \geq HMT4 \geq AIM > HMT3 > HMT2 > HMT1 \qquad (5.4.1)$$

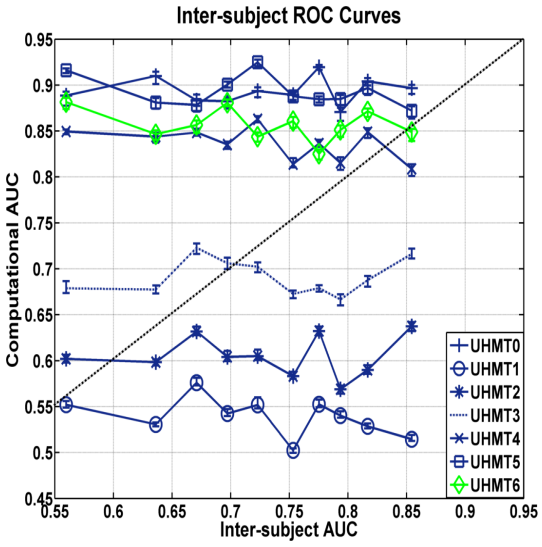
According to ROC curves in Figure 5.3, HMT with the smallest block (HMT5) and integrated HMT (HMT0) produce better performance than AIM, DIS do. It is noted that, AIM utilizes 30x30 windows, almost equivalent with (U/T/V) HMT1's window sizes. Moreover, AIM saliency maps is generated form data of overlapping blocks while non-overlapping blocks are used in MDIS. Utilization of such large windows and high-dimensional data in AIM requires training steps and samples preparation which require enormous computational effort. AIM uses continuously sliding windows and generates more detailed and smoother saliency maps than MDIS does with non-overlapping large blocks. For example, AIM performs better than HMT1 (32x32 blocks), HMT2 (16x16 blocks) and HMT3 (8x8 blocks). MDIS out-performs AIM if sizes of blocks are smaller than 8x8 .i.e HMT4 (4x4 blocks), HMT5(5x5 blocks).

In comparisons of (U,T,V) HMTs, advantages of (T,V)HMT over UHMT can be noticed on different scales. According to Figures 5.3a,5.3c,5.3e, ROC curves of (T,V)HMT move closer toward the left-top corner than UHMT does. It means (T,V)HMT more successfully detect meaningful points than UHTM does in region of low fall-out rate (FPR). Therefore, (T,V)HMT provides better binary classifier in term of robustness. It is due to the fact that THMT, VHMT require training steps on image data; meanwhile UHMT model just uses predefined and general parameters. Furthermore, VHMT has slightly better performance than THMT in most block sizes. The slight increment in terms of ROC curves is due to the fact VHMT is a more comprehensive model than THMT when processing texture features [146].

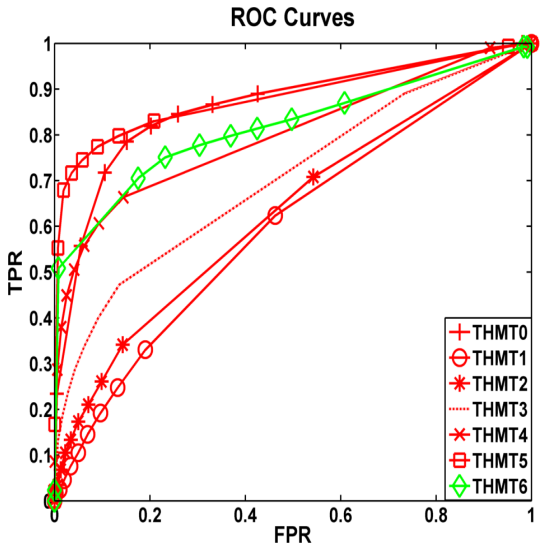
Generalized ROC evaluates performances of saliency maps with different thresholds, according to ground-truth eye-fixation data of several human test subjects. However, it does not distinguish eye-fixation data from each subject but treats the data collectively. Therefore, it certainly loses important aspects of eye-tracking data from multiple subjects such as diversity of individual responses over various types of scenes shown during experiments. To assess limitation of ROC evaluation, this chapter follows suggestion of Harel [49] by applying the ISROC evaluation method , as described in Sub-Section 2.3.2 . The correspondent simulation generates ISROC graphs 5.3b,5.3d, and 5.3f for MDIS with (UHMT) universal, (THMT) trained and (VHMT) vector-based configuration. Generally, all saliency methods MDIS, DIS and AIM have consistent performances across different complexity of scenes. They could produce meaningful saliency



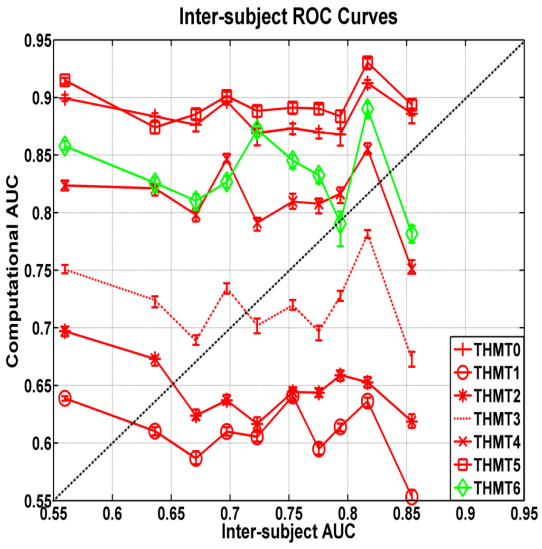
(a) UHMT - MDIS - ROC



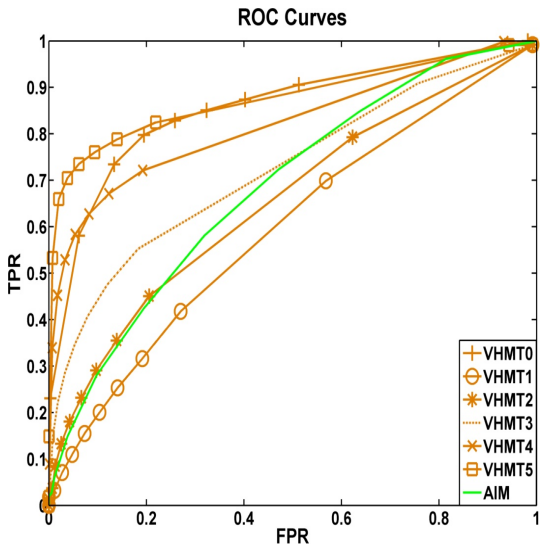
(b) UHMT - MDIS - ISROC



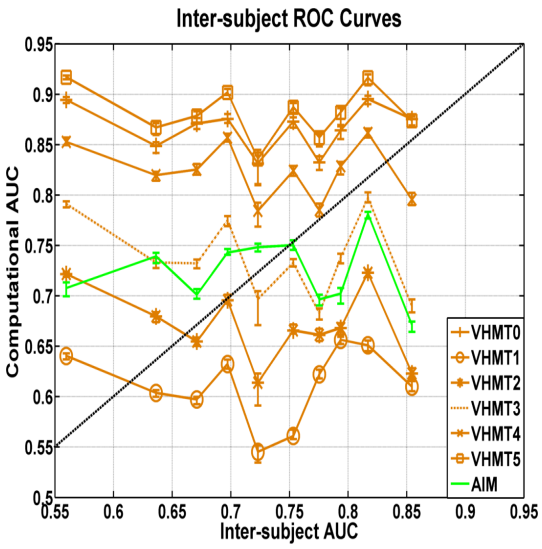
(c) THMT - MDIS - ROC



(d) THMT - MDIS - ISROC



(e) VHMT - MDIS - ROC



(f) VHMT - MDIS - ISROC

Figure 5.3: (U/T/V)HMT - MDIS - (IS)ROC

maps in complex cases where inter-subject scores are low, human subjects are struggling to find common salient points. However, these computational saliency maps are over-complex in simple scenes with high inter-subject ROC scores. In these cases, all subjects focus on a few locations on testing scenes, while the proposed saliency methods still detect other regions of interest beside main objects. Therefore, the proposed computational approaches are less efficient than human beings in such situations.

Similar to ROC, ISROC curves help to compare MDIS with different (U/T/V) HMT configurations on multiple levels (1-5). Observation of Figure 5.3 shows advantages of MDIS with different HMTs when smaller block sizes are chosen as ISROC curves of MDIS with (U/T/V) HMT5 are higher than (U/T/V) HMT4, etc. In accordance with ISROC curves in Figure 5.3, the reference methods AIM performance, the solid green line in Figure 5.3f, are generally better than HMT1 (32x32 blocks) and HMT2 (16x16 blocks), equivalent with HMT3 (8x8 blocks), but worse than HMT0, HMT4 (4x4 blocks), and HMT5 (2x2 blocks). Meanwhile, DIS, the solid green line in Figures 5.3d and 5.3d, is better than HMT1-3, almost equivalent to HMT4 and worse than HMT0 and HMT5. The ranking orders of HMT#, AIM and DIS in terms of ISROC are almost similar the rank in terms of ROC, as shown in the Equation 5.4.1. HMT0, HMT4 and HMT5 have better performance than human subjects in most of samples; their ISROC curves are above the black boundary set by human subjects for most of the plots. There is an interesting observation about HMT0 - the integrated map, and HMT5 - 2x2 blocks that HMT5 is almost equivalent with or sometimes better than the integrated method HMT0. Perhaps, more complex rules for integrating saliency maps need developing to efficiently integrate attention maps of different scales. Graphical curves relatively describe a general trend that MDIS with different HMTs out-perform the referenced AIM and DIS maps; however, it does not specify how much better the proposed methods are. Therefore, numerical results of performances need to be presented and analysed.

Table 5.1: Comparisons of Evaluation Results for MDIS, DIS and AIM

Observations	LCC	NSS	AUC	TIME
MDIS (THMT0)	**0.02382**	**0.48019**	**0.88357**	2.32734
DIS (UHMT6)	0.00553	0.46003	0.85658	0.88706
AIM	0.01576	0.12378	0.72400	50.41714

Results of numerical evaluation from MDIS, pseudo-DIS and AIM saliency approaches are shown in Tables 5.2, 5.3 and 5.4. The Table 5.3 highlights LCC, NSS and AUC of THMT0 which is the best configuration for MDIS method. Meanwhile the best result

Table 5.2: UHMT - MDIS - DATA

Observations	LCC	NSS	AUC	TIME (s)
UHMT0	0.01434	0.21811	**0.89392**	0.39617
UHMT1	-0.00269	0.19772	0.53862	0.39617
UHMT2	0.01294	0.27819	0.60520	0.39617
UHMT3	0.01349	0.32868	0.69065	0.39617
UHMT4	**0.01604**	**0.42419**	0.83615	0.39617
UHMT5	0.00548	0.13273	**0.89234**	0.39706
UHMT6 (DIS)	0.00553	0.46003	0.85658	0.88706

Table 5.3: THMT - MDIS - DATA

Observations	LCC	NSS	AUC	TIME (s)
THMT0	**0.02382**	**0.48019**	**0.88357**	2.32734
THMT1	**0.02582**	0.38096	0.60922	2.32734
THMT2	0.01156	0.31855	0.64633	2.32726
THMT3	0.01604	0.32491	0.71972.	2.32726
THMT4	0.01143	0.29662	0.81192	2.32726
THMT5	0.00512	0.36932	**0.89532**	2.32726
THMT6 (DIS)	-0.2673	0.13989	0.83324	7.28872

Table 5.4: VHMT - MDIS - DATA

Observations	LCC	NSS	AUC	TIME (s)
VHMT0	0.01697	**0.44170**	**0.86606**	2.84212
VHMT1	0.01693	0.38387	0.61187	2.84212
VHMT2	**0.02044**	0.38777	0.67060	2.84212
VHMT3	0.01430	0.38882	0.73682	2.84212
VHMT4	0.00946	0.36761	0.82329	2.84212
VHMT5	-0.00125	0.39580	**0.88160**	2.84212
AIM	0.01576	0.12378	0.72400	50.41714

of pseudo-DIS is from the highlighted UHMT6 row in Table 5.2. Together with AIM's evaluation results as highlighted in Table 5.4, best results of MDIS, DIS and AIM are arranged in Table 5.1. All three numerical results (LCC, NSS and AUC) in this table confirm the general trend, shown in ROC and ISROC graphs, that MDIS performs better than both (pseudo) DIS and AIM. Detailed analysis of the numerical evaluation results in the Tables 5.2, 5.3 and 5.4 is presented in the rest of this Sub-Section.

Computational loads are the first measure to be analysed and compared among saliency approaches. In the TIME rows of Tables 5.2,5.3 and 5.4, necessary processing time for each method or each mode is presented. Generally, computational loads, proportional to processing time, of all modes in either UHMT or THMT row is almost similar since the parameters of full-depth Hidden Markov Tree need estimating before computation of saliency values. In comparison of (U/T/V) HMTs (1-5), UHMT is faster than (T/V)HMT as UHMT uses predefined parameters instead of learning HMT parameters from each image. When comparing (U/T/V) HMT modes of MDIS with AIM, the proposed methods are much faster than AIM. The well-known AIM directly estimates self information from high-dimensional by ICA algorithm while MDIS statistically models two hidden states: "large" "small" states in sparse and structural features with efficient and fast inference algorithms. Computational load or processing time of the mentioned AIM and proposed MDIS with different modes can be seen in Figure 5.4d. This figure also shows comparison of computational times between MDIS, (U/T/V) HMT (1-5), and pseudo-DIS, (U/T) HMT6 in which MDIS is at least twice faster since DIS needs to estimate likelihood for dense distributions of pixel values.

Though MDIS with different HMTs significantly reduces computational load for computation of information-based saliency, more verifications are necessary for their performances in terms of accuracy. Evaluation of three modes (U/T/V)HMT separately are compared against AIM and DIS in terms of three numerical tools LCC, NSS, AUC together, Figure 5.4a, 5.4b, 5.4c. Then all modes of HMTs are summarized in three plots (the top row of Figure 5.5) in the following order NSS, LCC, AUC from left to right. Especially in Figure 5.5, simulation modes of the same block size are placed next to each other; for example, (U/T/V) HMT0 sit next to each other, so do (U/T/V) HMT1 and etc. It is intentionally arranged in that way to compare performances of different simulation modes in the same scale level. In Tables 5.2 and 5.3, **maximum** and `minimum` values among HMT modes of MDIS are notified by their corresponding text styles. In Figures 5.4a, 5.4b, 5.4c, extreme (maximum or minimum) values are also specially marked. For example, maximum values have big solid markers while big ones without any textures represent minimum points and marks with "brick-wall" textures are for

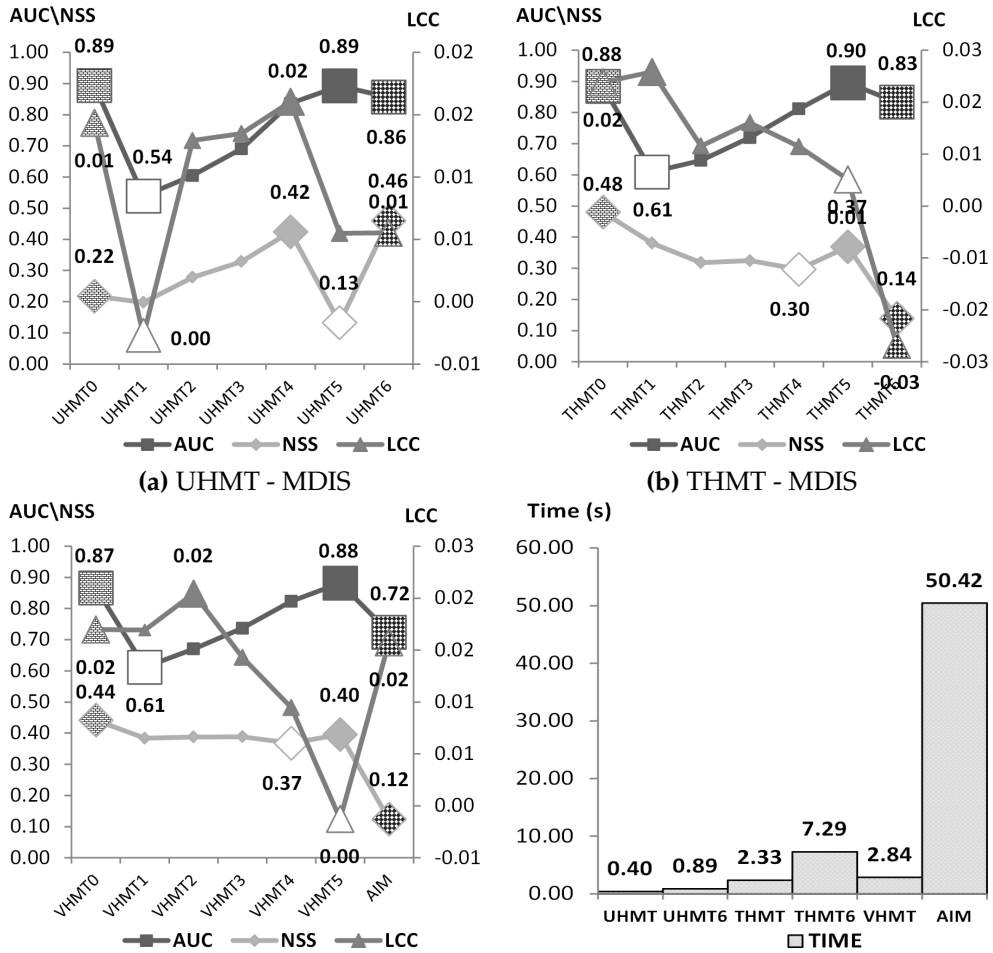


Figure 5.4: Performances of (U/T/V)HMT-MDISin AUC,NSS,LCC,TIME

integrated HMT methods. Especially, AIM and HMT6 (DIS) have big markers with distinguishing big cross-board texture while integrated saliency modes of (U/T/V)HMT0 have small cross-board textures. These special markers help highlight interesting comparisons of the proposed MDIS against AIM and DIS. The same marking policy is applied for data representation in Figure 5.5. Meanwhile, each line in this figure has an arrow head for showing trends of experimental data (increasing/decreasing) when simulation modes are changed across U,T, or V configurations of HMT for each scale level.

According to Figures 5.4d and 5.5, UHMT modes require very little effort in saliency computation. Obviously, that fact raises a question about its accuracy of centre and surround classifier as well as synthesized saliency maps. According to simulation data in the Table 5.2 with highlighted extrema, UHMT performs pretty well against AIM/DIS in all three measurements LCC, NSS and AUC. For example, MDIS with UHMT4 mode (4x4 square blocks) surpasses AIM in all measurements. It confirms validity and

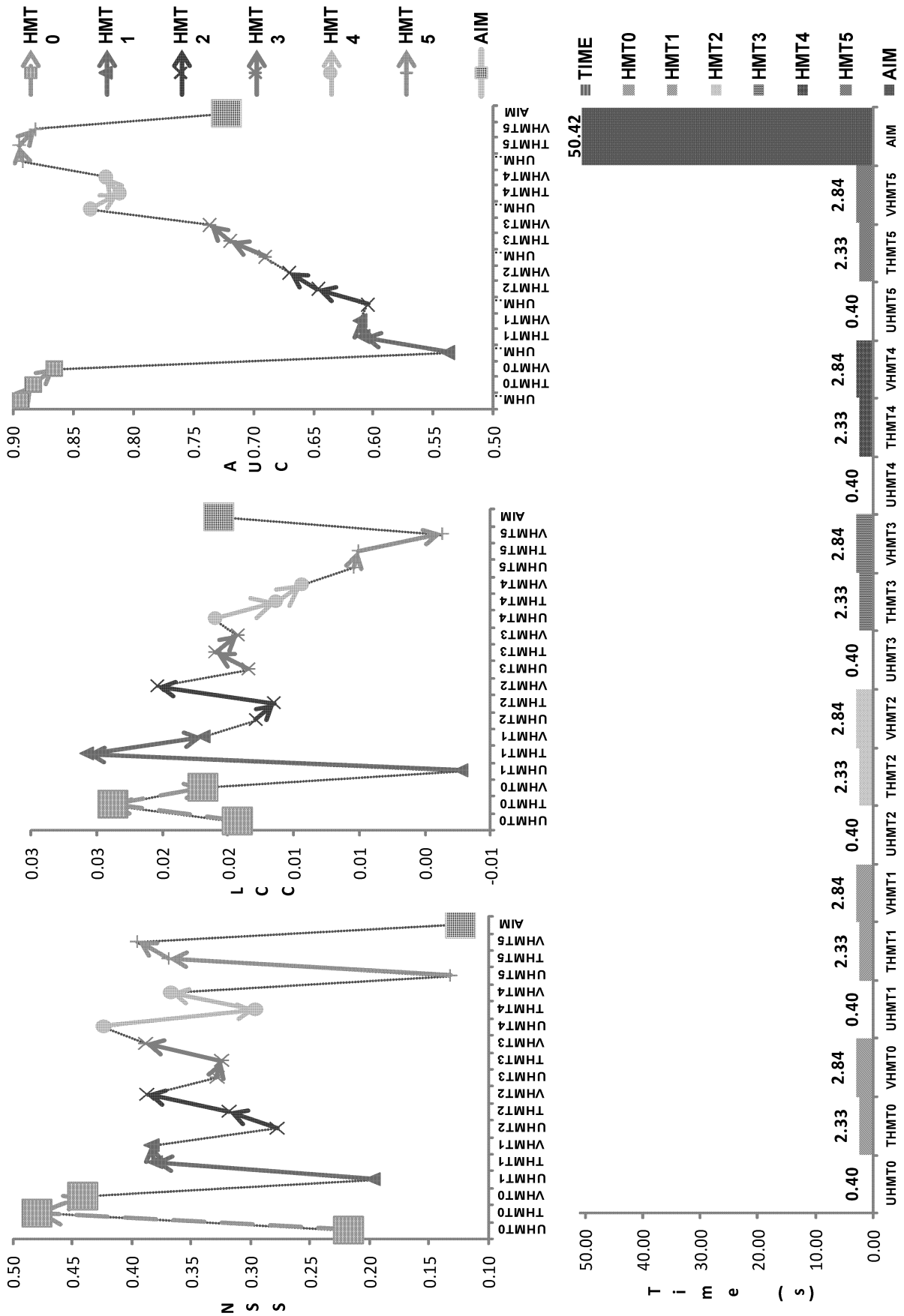


Figure 5.5: Summary of all MDIS against AIM/DIS

efficiency of MDIS in the information-based saliency map research field. When performances of different UHMT-MDIS modes are considered, UHMT4 with 4x4 squares have the most consistent evaluation among UHMT(1-5) with maximum LCC and NSS and the second best AUC value. UHMT0-MDIS, integration of saliency values across scales, does not have better performance than other UHMTs except for AUC level. The overall results show consitency of deploying HMT with predefined universal parameters for MDIS which requires no training effort for adapting the model into multi-scale statistical structures.

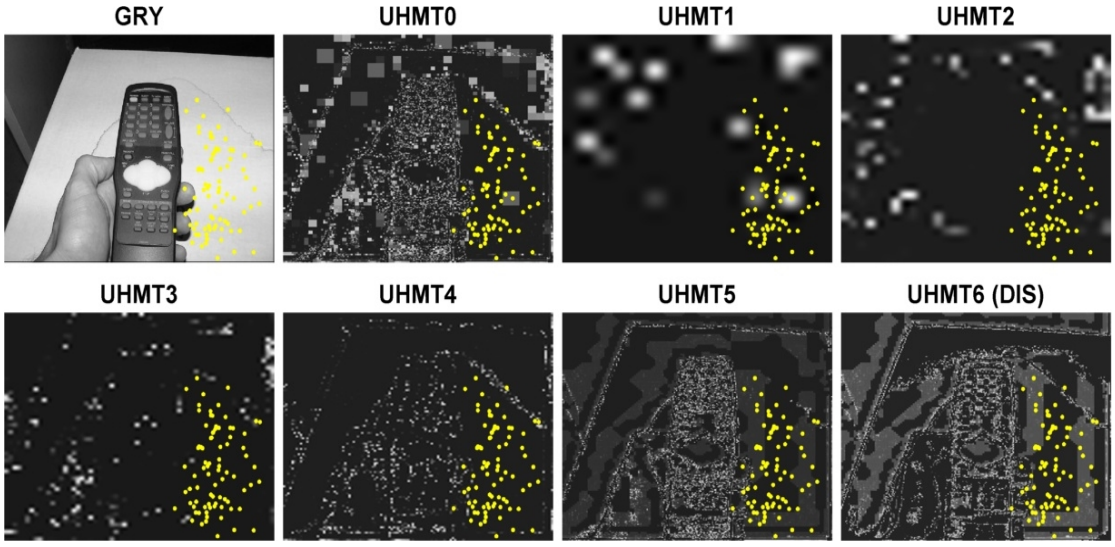
Different from UHMT mode, training stages are included in the simulation of MDIS with (T/V) HMT mode. With additional adaptivity, (T/V)HMT improves the saliency evaluation and produce more consistent results than UHMT does. It is shown through evaluation data in the Tables 5.3,5.4 and corresponding plots in the Figures 5.4b,5.4c. In these tables, all **maximum** values locate at the THMT0 column, THMT0-MDIS overperforms AIM and DIS (THMT6) in all evaluating schemes which confirms advantages of MDIS. Furthermore, effectiveness of training stages is clearly shown when comparing THMT0 against UHMT0. Though AUC of THMT0 is smaller than that of UHMT0, other evaluations of THMT0 are better than their counterparts in both NSS and LCC schemes. This confirms usefulness of training Hidden Markov Tree models for each sample image. In addition, the Figure 5.4b,5.4c shows supremacy of (T,V) HMT0 modes, the across-scale integration mode of MDIS over other singular saliency maps at different dyadic scales in any measurement. It is noted that LCC of THMT0 mode is a bit smaller than LCC of THMT1 mode; however, this small difference can be safely ignored. Comparison of (U/T/V)HMT-MDIS mode-by-mode between data in the Table 5.4a,5.4b,5.4c are shown in Figure 5.5. Accordingly, there are slight improvement in performance of (T/V)HMT1,(T/V)HMT2 over UHMT1, UHMT2, equivalence between (T/V) HMT3 and UHMT3, and a reverse trend that UHMT4-5 are comparable or slightly better than THMT4, THMT5. It seems that training processes are more important when big classification windows are used. Meanwhile universal approaches of HMT work pretty well if dyadic squares get smaller. Two possible reasons for this observation are statistical natures of dyadic squares and characteristics of training processes. A bigger square has richer joint-distribution of features; therefore, UHMT with fixed parameters can not marginally approximate that distribution well. However, parameters of (T,V)HMT models can be learnt from analysing images; it results in significant improvement of saliency maps quality. While smaller sub-squares are less statistically distinguishing, they are successfully modelled by universal parameters of HMT. In these cases, training processes might become redundant since UHMT would perform as well as THMT would do.

5.4.2 Qualitative Evaluation

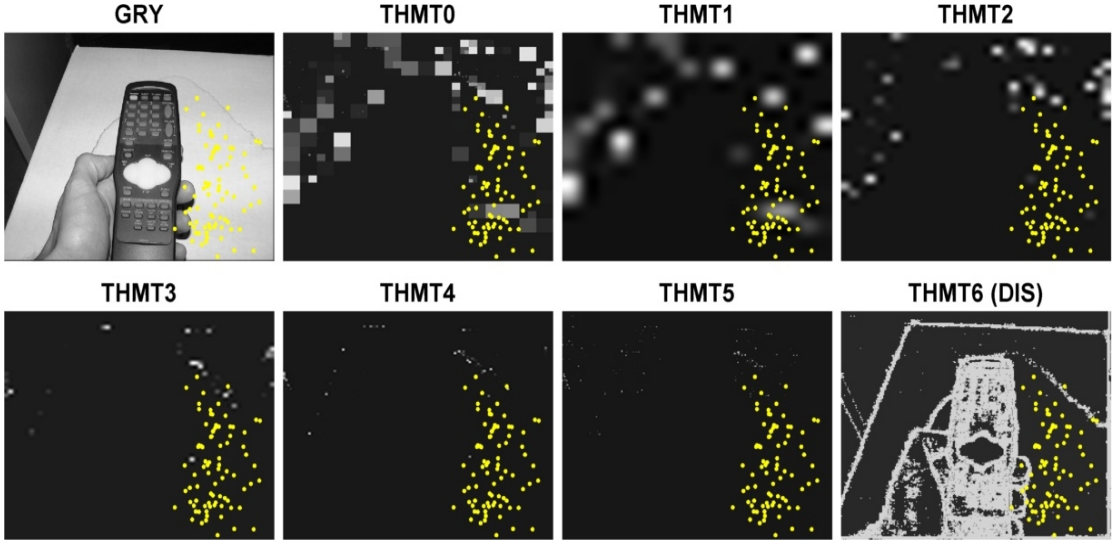
In this section, saliency maps are analysed qualitatively or visually. In addition, ground-truth eye-fixation points are marked on both gray-scale images and their saliency maps by yellow points. This analysis helps identify (i) on which image contexts (U/T/V) HMT-MDIS work well, (ii) how scale parameters affect formation of saliency maps, and (iii) how MDIS in general is compared with AIM/DIS.

In Figures 5.6 and 5.7, the first example with big central objects show an example of good (U/V)HMT performance but bad THMT performance. All scale levels of THMT suppress features of the most obvious object in the image centre. Meanwhile, (U/V)HMT4 and (U/V)HMT5 capture significant features of those objects; therefore, (U/V)HMT0 has much better saliency map than THMT0. In this case, the best saliency map of MDIS approach, (U/V)HMT0, is reasonably competitive against AIM/DIS. Despite different configurations, UHMT-MDIS and VHMT-MDIS have quite similar results for simple scenes with few central objects. Figures 5.6a, 5.7a and 5.6c,5.7c show differences between generated saliency maps of UHMT and VHMT across six different modes [0-5] especially mode 5. This mode utilizes the 2x2 blocks; therefore, it can create saliency maps with great details about discrepant regions in (U/V)HMT modes. Generally, UHMT tends to include more irrelevant areas than VHMT does; while VHMT only focuses on regions richer of edges and textures. It is coherent with quantitative comparisons between UHMT and VHMT in the tables 5.2, 5.4. However, there exists exceptions like the case in Figure 5.7 where VHMT misses edges and texture of foreground objects but wrongly focus on small black circles on the background. Though universal mode does not have on-line training, its predefined parameters are learnt by off-line learning from several natural images. Therefore, it robustly performs in general scenes which others fail to operate well. Meanwhile, saliency maps generated by VHMT-MDIS are much attracted to small black details on background due to rotation-invariant nature of the vector-HMT model [146].

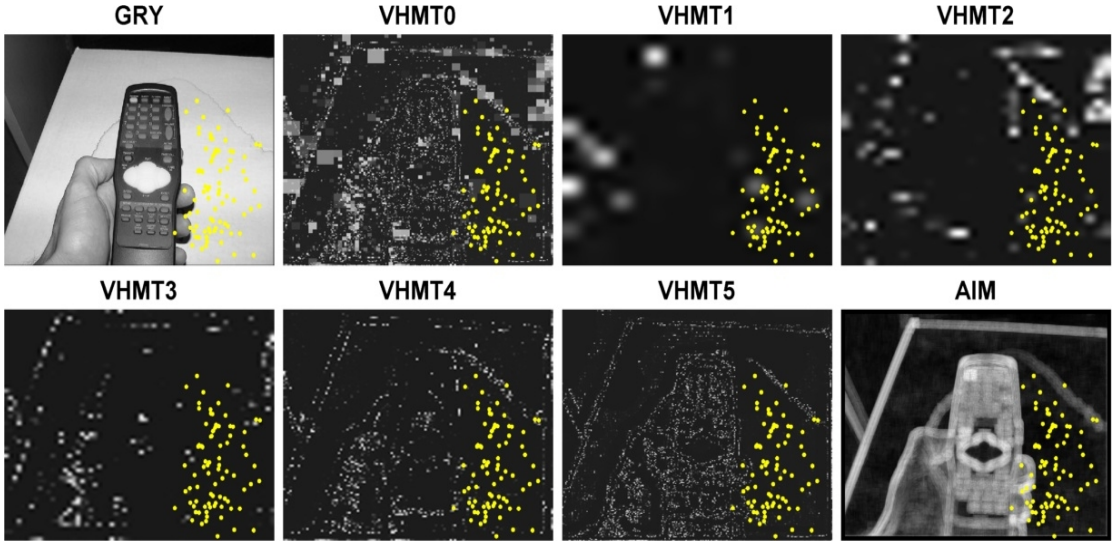
In contrast to the previous examples, Figures 5.8, 5.9 show an opposite case in general outdoor scenes for which (T/V)HMT produces more reasonable saliency maps. While UHMT0 map covers the whole region of sky despite no interesting features, (T/V)HMT0 correctly focuses on interesting but scatter features on the scene. Similarly, (V/T)HMT does extract more meaningful features than UHMT does (see the figures 5.8,5.9). In addition, the best saliency map THMT0 or THMT5 highlights more discriminant features than AIM/DIS saliency map. Noteworthy, there are significant differences between VHMT1 and THMT1 saliency maps. While THMT1 over-emphasizes edge points and lines between appeared textures such as trees and sky, the VHMT1



(a) UHMT

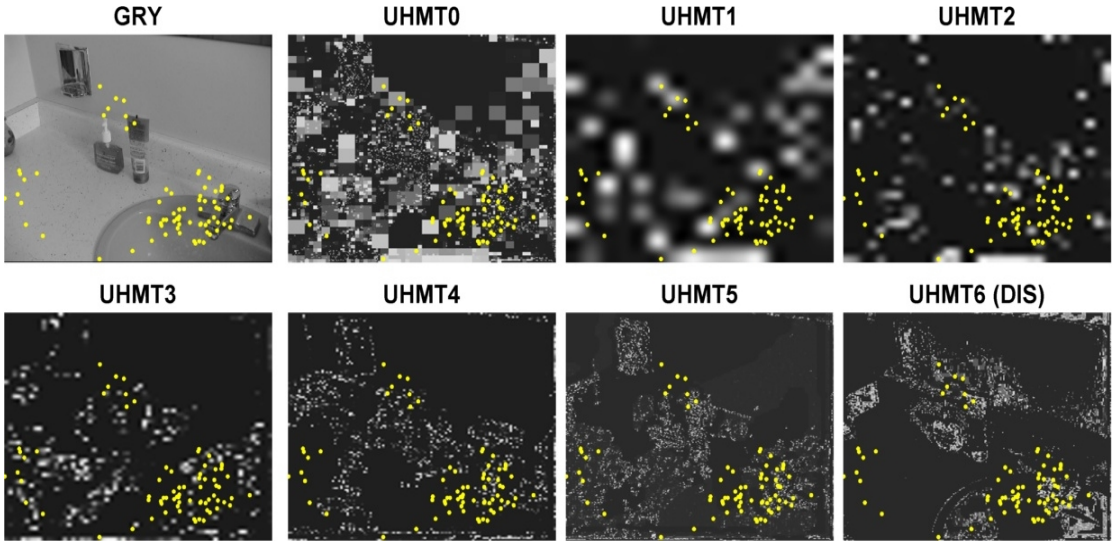


(b) THMT

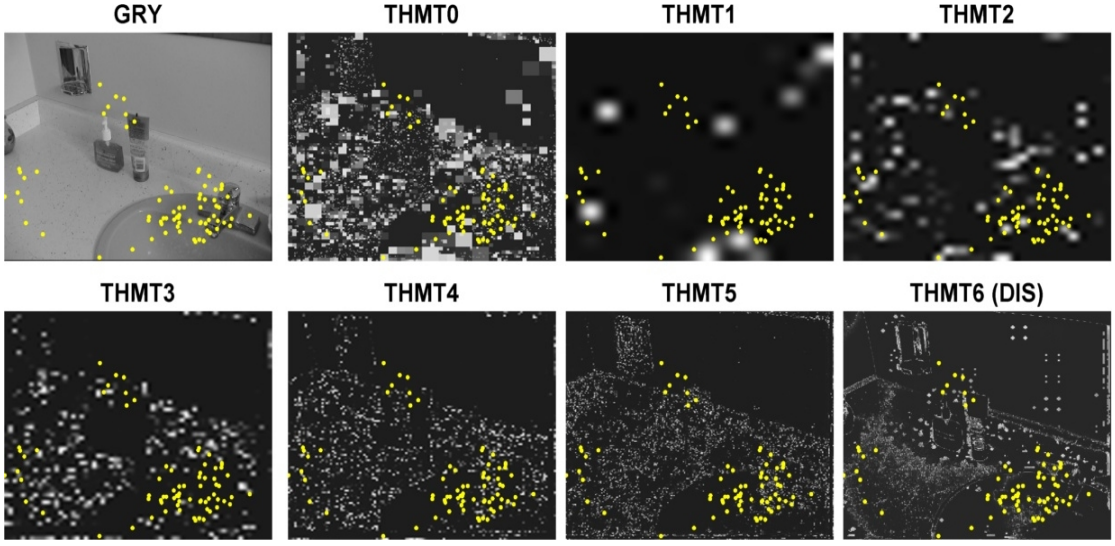


(c) VHMT

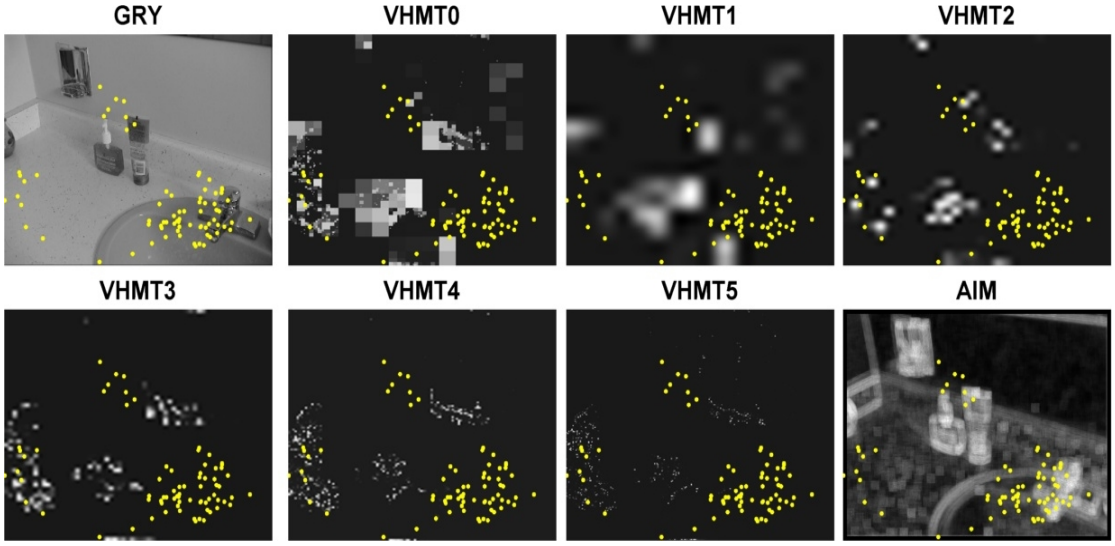
Figure 5.6: Saliency Maps 1A



(a) UHMT

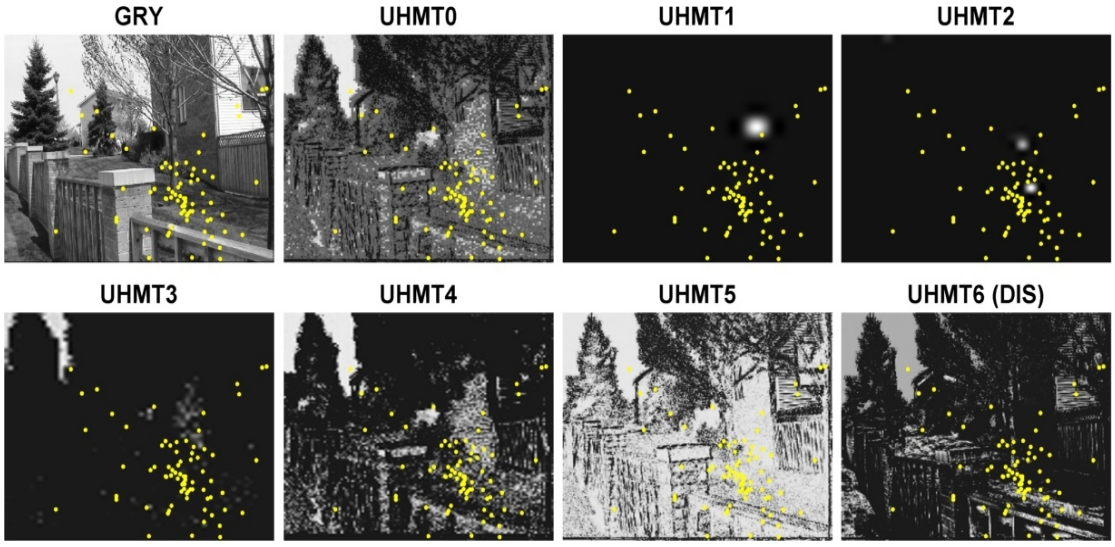


(b) THMT

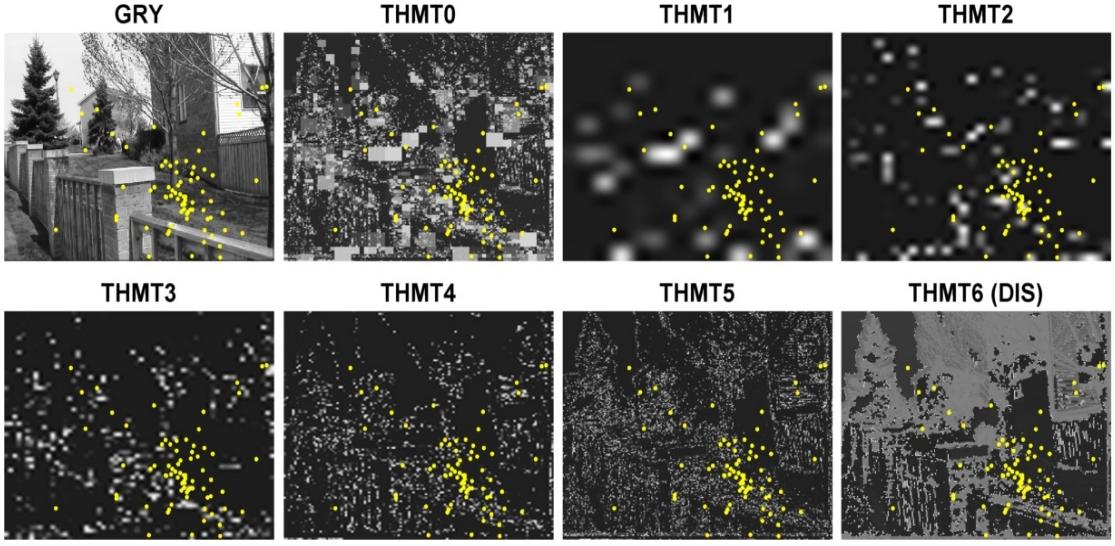


(c) VHMT

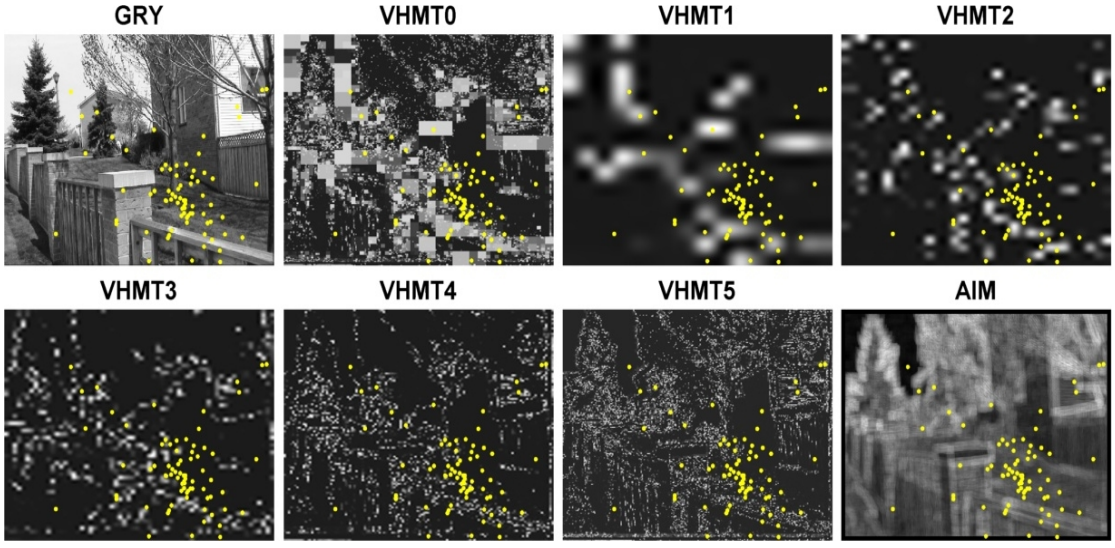
Figure 5.7: Saliency Maps 1B



(a) UHMT

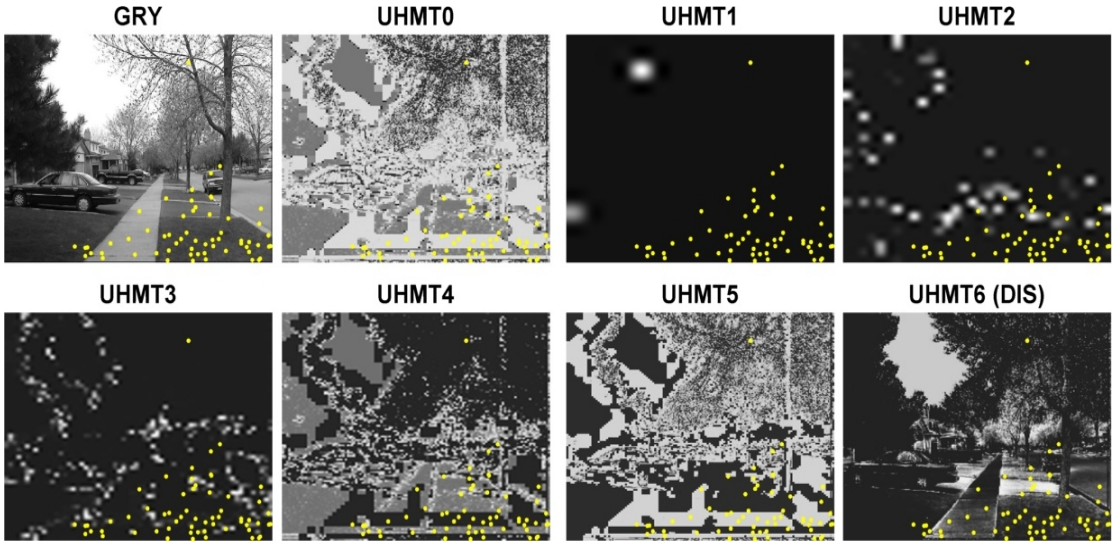


(b) THMT

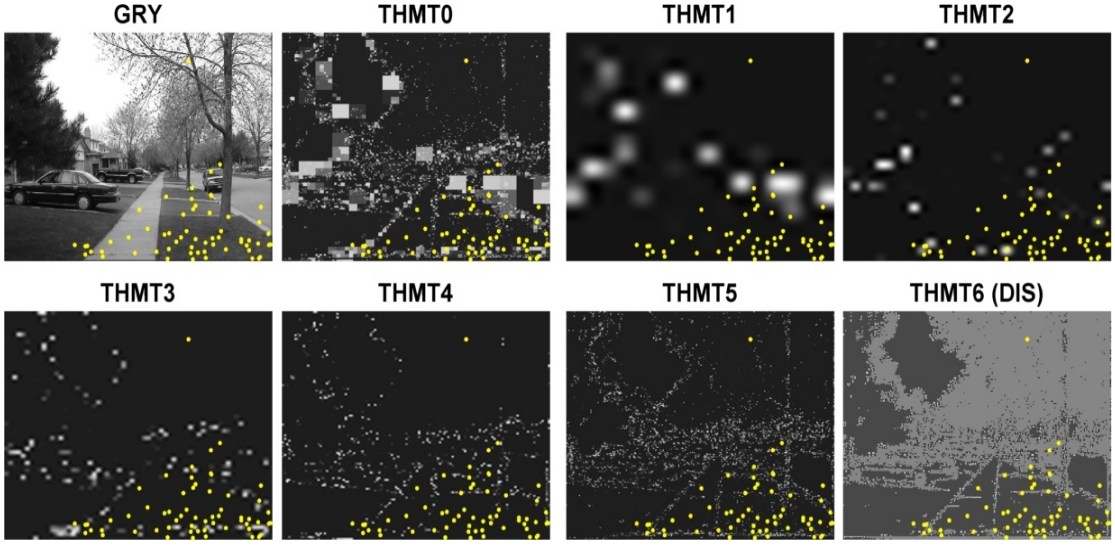


(c) VHMT

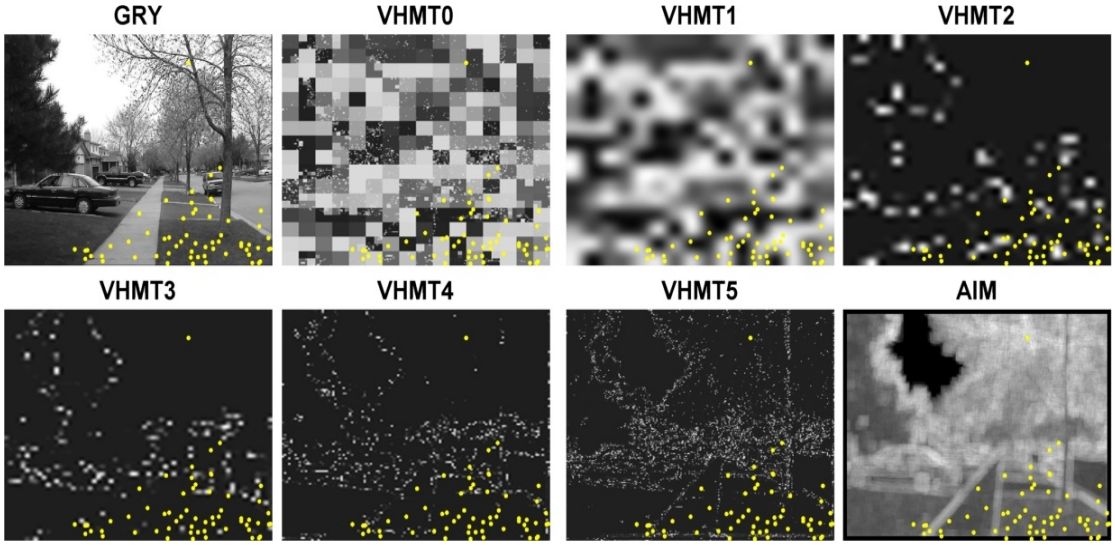
Figure 5.8: Saliency Maps 2A



(a) UHMT

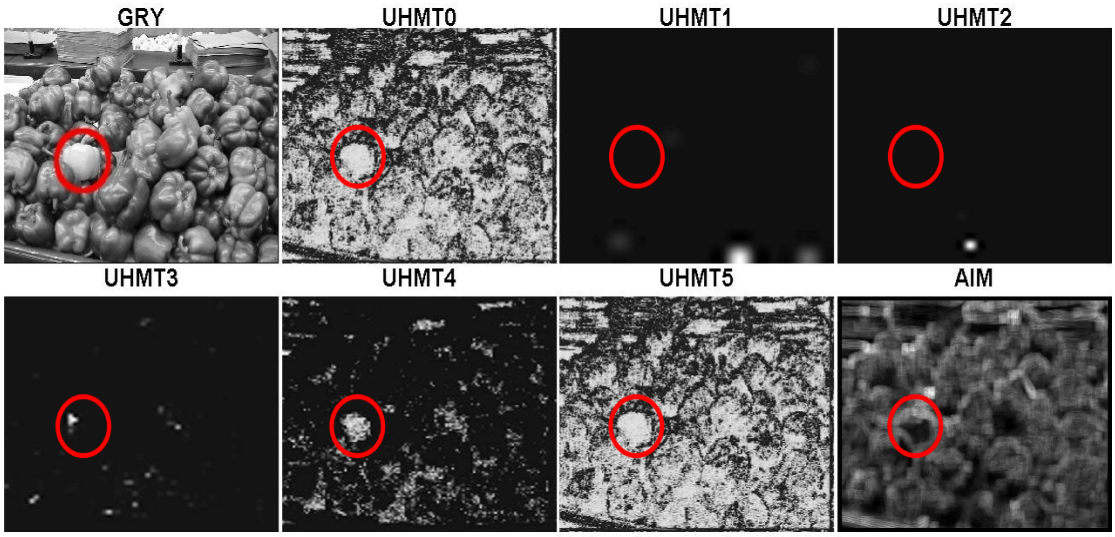


(b) THMT

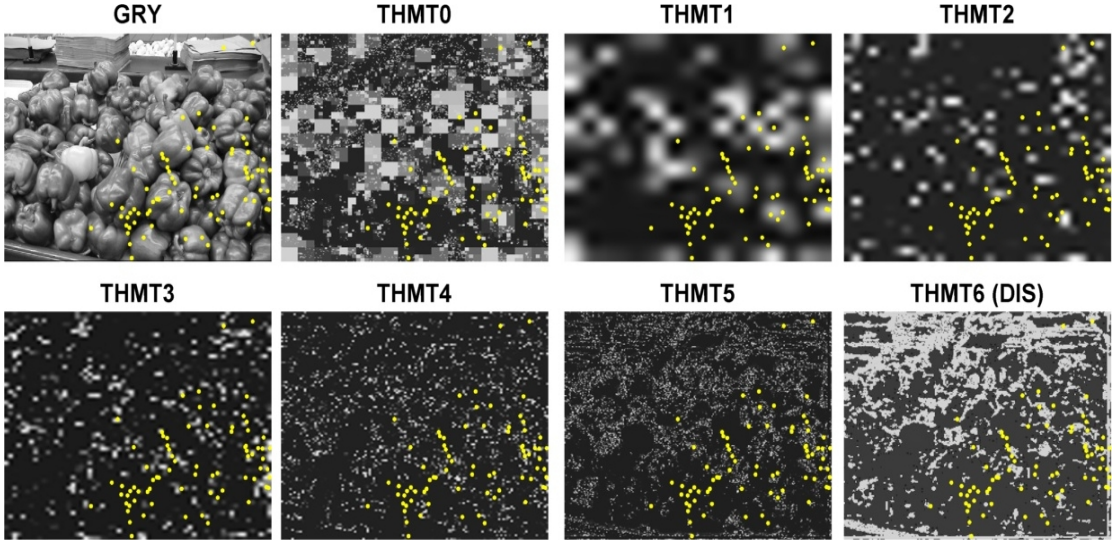


(c) VHMT

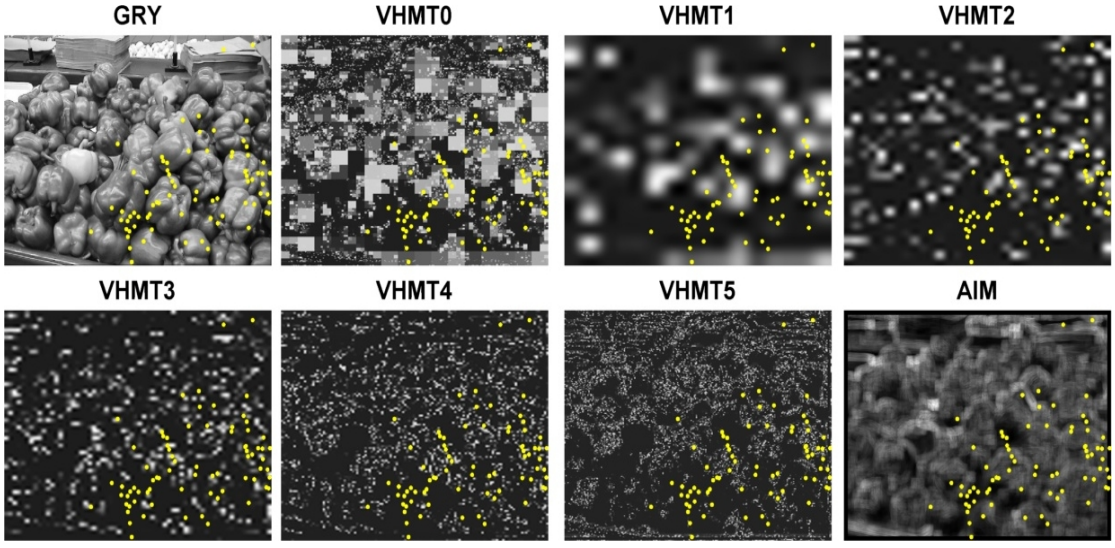
Figure 5.9: Saliency Maps 2B



(a) UHMT



(b) THMT



(c) VHMT

Figure 5.10: Saliency Maps 3A

is ridged with highlighted regions and does not hint any standing-out areas. Again this discrepancy in performance may be due to nature of orientation selection in each mode (T/V) as THMT favours only features of horizontal, vertical, diagonal directions; while VHMT mode, a rotation invariant scheme, does not have any oriental differentiation. The third example is chosen such that complex scenes are presented to the saliency methods. In the Figure 5.10, there are several fruits on the shelf; it is considerably complicated due to richness of edges, textures, as well as colour. In general, all (U/T/V)HMT-MDIS and AIM/DIS only partially succeed in detecting saliency regions from these images since none of them successfully highlight the distinguished fruit on the shelf (the fruit inside a red circle, as shown in Figure 5.10) . Though most MDISs for variety of scale levels, do not explicitly detect that fruit, UHMT3 and UHMT4 salency maps are able to highlight the location of that fruit (see UHMT3 an UHMT4 saliency maps, the figure 5.10)). The sample matches with the fact that UHMT4 data in the Table 5.2 has extremely good performance in all evaluation schemes. Surprisingly, there are some cases when appropriate choices of scales and parameters of predefined HMT models can over-perform all trained HMT models. The interesting examples of the Figure 5.10 opens another research direction about how HMT model can be learned optimally; however, it is the question of further research.

5.5 CONCLUSION

Multiple Discriminant Saliency (MDIS) aims to extract information from non-linear statistical dependencies between wavelet across adjacent dyadic scales. As saliency concept of MDIS is motivated by decision theory, contrast-based centre-surround operation of saliency in biological human visual system is formulated as minimizing error in classification of central and surrounding features. In this chapter, MDIS adapts the decision-theory like saliency concept in multiple scales by replacing these central and surrounding features with finer and coarser features of consecutive scales. Minimization of error rate or optimization of saliency values bases on maximum-a-posterior probability mechanism across multiple dyadic scales from Wavelet Hidden Markov Tree (WHMT). This optimization approach has been utilised in image segmentation, a binary operator; meanwhile, this chapter utilises this similar method to optimize decision-theory saliency in scale-space, a derivative operator. Performances of MDIS are evaluated against An Information Maximization (AIM) and Discriminant Saliency (DIS) approaches on Niel Bruce's database with eye-tracking coordinates as ground-truth. Furthermore, standard quantitative tools such as AUC, NSS, LCC, ROC, ISROC are utilised to ensure thorough comparison among considered saliency methods. Be-

sides, qualitative comparisons are done by visual comparison between maps, generated from chosen image samples by these saliency approaches. These visual evaluation highlights competitiveness of MDIS against AIM and DIS; however, they also reveals MDIS's limitations, e.g. missing salient areas on few complex scenes. Despite common disadvantages of information-based saliency techniques, MDIS brings discriminant saliency concepts into scale space and estimate saliency value from non-linear statistical relation across dyadic scales with established statistical optimization technique.

CHAPTER 6

Conclusion and Future Works

Information-based saliency models attention as problems of information estimation assumed that the more informative a location is, the more salient it is to visual system. However, information estimation of high-dimensional data like images and videos suffers from the curse of dimensionality. This curse makes information estimation a difficult and slow process. These limitations motivate us to improve and develop alternative information-based saliency techniques that are biologically plausible and fast for computer vision applications. In pursuit of robust and fast saliency techniques, three solutions ENT, WSS and MDIS have been introduced in Chapters 3, 4 and 5. Though all three methods aim to tackle the curse of dimensionality in information-based saliency approach, the latter methods are designed more sophisticatedly than the previous methods. Their relations and contributions are summarized in following paragraphs and Figure 6.1.

Entropy-based Saliency Method (ENT) The first proposal ENT is applying fast kd-tree entropy estimation method (KDPEE) [21] on Medium Sub-band Frequency filtered (MSF) to estimate statistical differences between data from central and surrounding regions. One of its rationales are reducing effects of high dimensionality by filtering only attention-related spectrum of data. The other rationales behind ENT are implementing a fast, non-parameters KDPEE method instead of building histogram for information computation and modelling saliency as non-linear centre-surround contrast . Then, this efficient approach is extended into spatio-temporal domain to process input videos rather than just still images. Moreover, it is also used for evaluating drivers' visual attention on roads.

Wavelet-based Scale Saliency (WSS) While ENT reduces data dimension in frequency domain but estimates information in spatial domain, the second proposal WSS utilizes

both spatial and spectral data in wavelet-based descriptors. The advantage of this approach in WSS is using sub-band energy descriptors and their distributions to evaluate information as saliency value. Furthermore, the utilization of the energy descriptors allows possible integration with scale saliency framework [22] by an introduction of a dyadic-scale frame-work and a modified mathematical model. The contribution of WSS is providing a fast method of saliency estimation across multiple dyadic scale; however, WSS, like ENT, only accounts for local features rather than global knowledge.

Multiscale Discriminant Saliency Method (MDIS) Despite being similar WSS method, the third proposal MDIS, using Hidden Markov Tree (HMT) framework, acquires alternative interpretation of non-linear centre-surround operation. Instead of direct measurement in informative discrepancies, MDIS adapts probabilistic solution in discriminating features of a central class from a surrounding class or vice versa. This probabilistic estimation of discriminant power can be efficiently done for multiple scales by Wavelet Hidden Markov Tree and its statistical inferencing process. In addition, the MDIS-HMT framework allows integration of prior knowledge such as specific statistical characteristics of natural images, which eliminates training stage for learning model parameters. The contributions of MDIS-HMT as well as other improvements of ENT and WSS are all aimed to deal with high dimensional nature of imaging data and reduce computational load but still guarantee accurate saliency values estimated by the proposed information-based methods.

Achieved Objectives In general, most of objectives, as mentioned in Sub-Section 1.3 of the Chapter 2, have been investigated and partly fulfilled in the above three main contributions as follows.

1. ENT method utilises conditional entropy and Kullback Leilber divergence as measurements of saliency.

2. MDIS formulates saliency centre-surrounding operator of saliency by decision theory by measuring how discriminative central regions are from surrounding regions.

3. Instead form image histogram and marginal distribution, high-order statistical information is estimated from Fast kd-tree entropy estimation (KDPEE) in ENT, energy distribution of wavelet descriptors in WSS, and Wavelet Hidden Markov tree (WHMT) in MDIS.

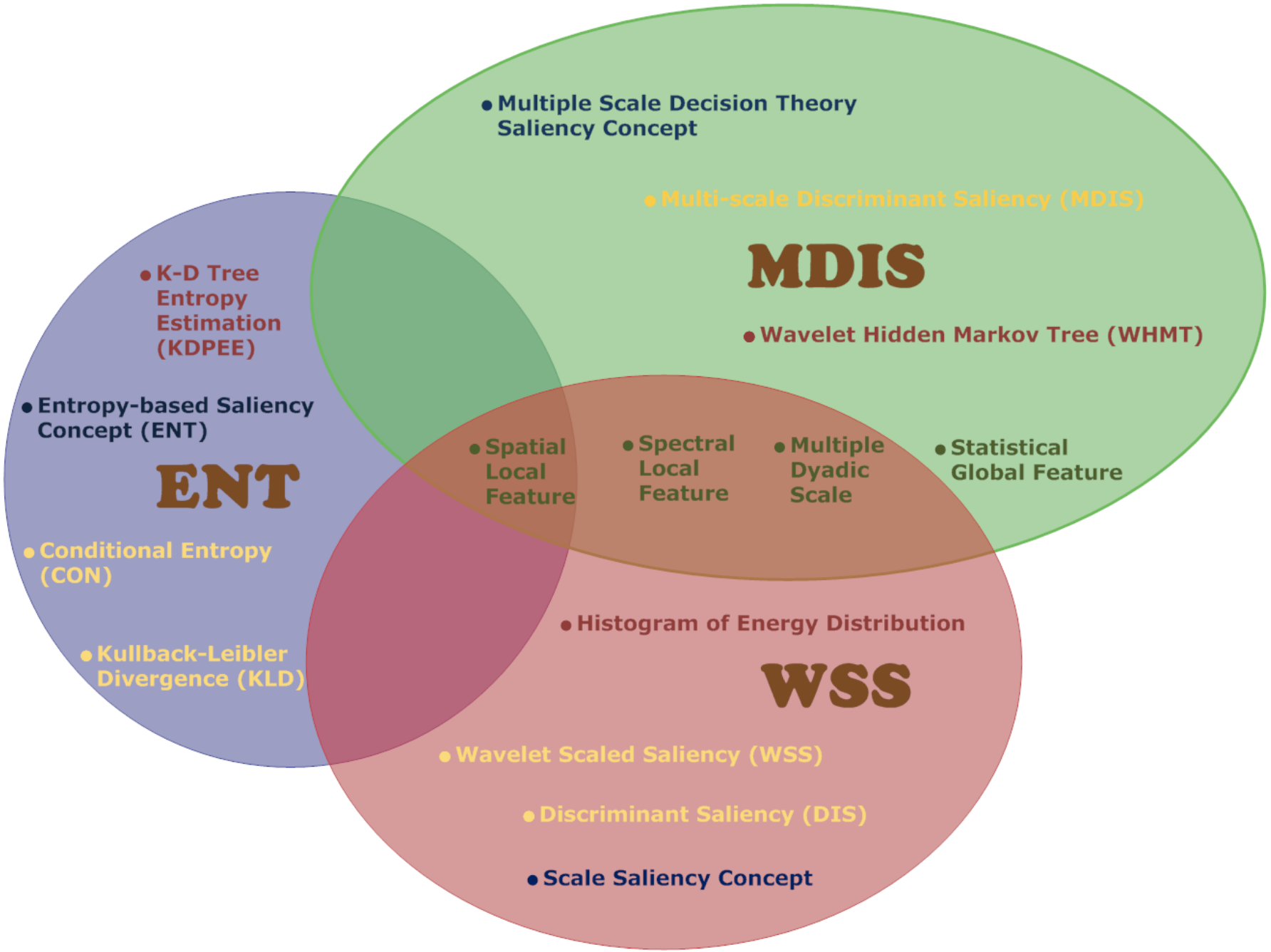


Figure 6.1: Relations and Overlapping between proposed methods ENT, WSS and MDIS

4. Strong theoretical background behind WSS and MDIS are well related to other established information-based saliency approaches like Scale Saliency [22], Bayesian Surprise Saliency [19], and Discriminant Saliency [92].
5. Information estimation has been done across dyadic scales in WSS and MDIS; in other words, these methods have considered multi-scale natures of visual attention. Moreover, prior knowledge about statistics of natural images, embedded in MDIS saliency, is shown to improve its performance.
6. ENT is extended and tested on both still images and videos.
7. All three proposals ENT, WSS and MDIS are shown to correlate well with ground-truth data, eye-fixation data, quantitatively.

Unachieved Objectives Despite much research effort, there are still many drawbacks in these three proposed information-based saliency methods

1. ENT suffers from bias in information estimation and its window sizes are limited by KDPEE method. In other words, it is difficult to extend ENT for multiple-dyadic scales.
2. WSS only focuses on local information across multiple scales does not include any prior-information or global information about natural images. Moreover, WSS only works for still images not video.
3. Similar to WSS, MDIS only work on still images rather than videos.

Future works Multiple dyadic scales and prior-information of natural images, two important characteristics of biological vision systems, have been integrated in proposed algorithms in this thesis. However, to achieve robust and fast biological plausible solutions for visual attention and saliency. several characteristics of biological systems needs to be considered and integrated into ENT, WSS or MDIS. For instances,

- Sparseness has widely been agreed as important aspect of spike train signals. [160]
- Predictive coding is proposed by Feldman and Friston. [96]

Furthermore, advancement in theoretical understanding of biological visual attention does as well push development of several smart solutions for computer vision problems since saliency techniques are often employed in feature extraction stages, initial steps of many applications. For examples,

- Medical image processing often encounters enormous high-resolution data. Doctors and therapists spend many hours to go through all information to check for any possible health-related threats to patients. With fast and robust information-based saliency techniques, automatically providing clues about interesting information within seconds, doctors only need to check specifically guided places instead of other insignificant details.
- Advanced Driving Assistance System (ADAS) needs to process data from several sensors in real-time. In other words, the system needs to be able to focus on information-rich area to process fast enough and help drivers make decisions. With fast and robust information-based sailency techniques, performances of ADAS can be improved in both accuracy and speed.

The information-based saliency would help boost quality of vision applications, detecting potential threats on time as well as reduce wasteful computational loads at uninterested area. Lists of computer vision applications, benefited from information-based saliency, are not limited in the above two fields since results of these information-based saliency methods can be integrated with any frameworks of computer vision systems.

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, p. 1254–1259, 1998. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=730558

[2] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, Jun. 2008, pp. 1 –8.

[3] F. Attneave, "Some informational aspects of visual perception." *Psychological review*, vol. 61, no. 3, pp. 183–193, 1954. [Online]. Available: http://psycnet.apa.org/journals/rev/61/3/183/

[4] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Advances in neural information processing systems*, vol. 18, p. 155, 2006. [Online]. Available: http://www.cse.yorku.ca/~tsotsos/Homepage%20of%20John%20K_files/NIPS2005_0081.pdf

[5] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," *2009 IEEE 12th International Conference on Computer Vision*, p. 2106–2113, Sep. 2009. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5459462

[6] Z. Azimifar, M. Amiri, P. Fieguth, and E. Jernigan, "Empirical study of wavelet domain image joint statistics and proposition of an efficient correlation map," *Journal of Mathematical Imaging and Vision*, vol. 43, no. 1, p. 24–38, 2012. [Online]. Available: http://www.springerlink.com/index/T78N34532587K013.pdf

[7] DFT, "Traffic signs and signals," http://www.dft.gov.uk/trafficsignsimages/UserGuide.php#_Toc204593704, Jan. 2010. [Online]. Available: http://www.dft.gov.uk/trafficsignsimages/UserGuide.php#_Toc204593704

[8] D. Walther, "Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics," Ph.D. dissertation, California Institute of Technology, 2006. [Online]. Available: http://thesis.library.caltech.edu/895/

[9] K. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain," *Journal of Physiology-Paris*, vol. 100, no. 1, p. 70–87, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092842570600060X

[10] M. I. Posner and S. Dehaene, "Attentional networks," *Trends in Neurosciences*, vol. 17, no. 2, pp. 75–79, 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0166223694900787

[11] D. H. Ballard, "Animate vision," *Artificial Intelligence*, vol. 48, no. 1, pp. 57–86, Feb. 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0004370291900804

[12] A. Yarbus, B. Haigh, and L. Rigss, *Eye movements and vision*. Plenum press New York, 1967, vol. 2. [Online]. Available: http://library.wur.nl/WebQuery/clc/340878

[13] U. Neisser, *Cognitive Psychology*. East Norwalk, CT, US: Appleton-Century-Crofts, 1967. [Online]. Available: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/049550629X

[14] W. Richards and L. Kaufman, ""Center-of-gravity" tendencies for fixations and flow patterns," *Perception & Psychophysics*, vol. 5, no. 2, pp. 81–84, Mar. 1969. [Online]. Available: http://link.springer.com/article/10.3758/BF03210526

[15] J. K. Tsotos, S. M. Culhane, W. Yan, K. Wai, and N. Davis, "Artificial intelligence modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 3702, no. 95, 1995.

[16] N. Bruce and J. Tsotsos, "An information theoretic model of saliency and visual search," *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, p. 171–183, 2007. [Online]. Available: http://www.springerlink.com/index/E983160657777854.pdf

[17] D. Gao and N. Vasconcelos, "Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics," *Neural Computation*, vol. 21, no. 1, p. 239–271, 2009. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/neco.2009.11-06-391

[18] N. D. Bruce, "Features that draw visual attention: an information theoretic perspective," *Neurocomputing*, vol. 65–66, pp. 125–133, Jun. 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231204004539

[19] L. Itti and P. Baldi, "Bayesian surprise attracts human attention." *Vision research*, vol. 49, no. 10, p. 1295–306, Jun. 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2782645&tool=pmcentrez&rendertype=abstract

[20] F. Urban, B. Follet, C. Chamaret, O. Le Meur, and T. Baccino, "Medium spatial frequencies, a strong predictor of salience," *Cognitive Computation*, vol. 3, no. 1, p. 37–47, 2011. [Online]. Available: http://www.springerlink.com/index/10.1007/s12559-010-9086-8

[21] D. Stowell and M. Plumbley, "Fast multidimensional entropy estimation by k-d partitioning," *IEEE Signal Processing Letters*, vol. 16, no. 6, p. 537–540, Jun. 2009. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4897236

[22] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, p. 83–105, 2001. [Online]. Available: http://www.springerlink.com/index/T45N2G8543574026.pdf

[23] G. Kootstra and L. R. B. Schomaker, "Prediction of human eye fixations using symmetry," in *Proceedings of the 31st Annual Conference of the Cognitive Science Society (CogSci09)*, 2009, p. 56–61. [Online]. Available: http://kth.diva-portal.org/smash/record.jsf?pid=diva2:454614

[24] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, p. 97–136, 1980. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0010028580900055

[25] J. Tsotsos, *A computational perspective on visual attention*, 2011. [Online]. Available: http://books.google.co.uk/books?hl=en&lr=&id=eT7dTizZfrYC&oi=fnd&pg=PP1&dq=A+Computational+Perspective+on+Visual+Attention+&ots=LCjjGHEKzY&sig=9zepqCdfhDV5U3F2hR1fUll9IDYhttp://books.google.com/books?hl=en&lr=&id=eT7dTizZfrYC&oi=fnd&pg=PP1&dq=A+Computational+Perspective+on+Visual+Attention&ots=LCjjGHEJE-&sig=CYmmi3FRSapWmSyeeKQu-zPHhW0

[26] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532 – 540, Apr. 1983.

[27] Tsotsos, "A 'complexity level'analysis of immediate vision," *International Journal of Computer Vision*, vol. 1, no. 4, p. 303–320, 1988. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=85505761752009200 3related:Y-PGcl_G3QsJ

[28] C. H. Anderson and D. C. Van Essen, "Shifter circuits: a computational strategy for dynamic aspects of visual processing." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 84, no. 17, p. 6297–301, Sep. 1987. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=299058&tool=pmcentrez&rendertype=abstract

[29] P. Burt, "Attention mechanisms for vision in a dynamic world," *ldots Recognition, 1988., 9th International Conference on*, 1988. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=28419

[30] K. Nakayama and G. G. Silverman, "Serial and parallel processing of visual feature conjunctions," *Nature*, vol. 320, no. 6059, p. 264–265, 1986. [Online]. Available: http://visionlab.harvard.edu/members/ken/Kenpapersforwebpage/043Nature86.pdf

[31] L. Uhr, "Layered "Recognition cone" networks that preprocess, classify, and describe," *IEEE Transactions on Computers*, vol. C-21, no. 7, pp. 758 – 768, Jul. 1972.

[32] S. Palmer, *Vision science: Photons to phenomenology*. MIT Press, 1999, vol. 4, no. 4. [Online]. Available: http://119.93.223.179/ScienceDirect/CognitiveSciences/04-04/sdarticle_012.pdf

[33] D. Marr and A. Vision, "A computational investigation into the human representation and processing of visual information," *WH San Francisco: Freeman and Company*, 1982. [Online]. Available: http://www.contrib.andrew.cmu.edu/~kk3n/80-300/marr2.pdf

[34] Z. Li, "A saliency map in primary visual cortex," *Trends in cognitive sciences*, vol. 6, no. 1, p. 9–16, 2002. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.8879&rep=rep1&type=pdf

[35] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry." *Hum Neurobiol*, vol. 4, no. 4, p. 219–27, 1985. [Online]. Available: http://papers.klab.caltech.edu/104/1/200.pdf

[36] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature*

reviews neuroscience, vol. 2, no. February, p. 1–11, 2001. [Online]. Available: http://papers.klab.caltech.edu/83/1/391.pdf

[37] J. Clark and N. Ferrier, "Modal control of an attentive vision system," *Proc. Int. Conf. Computer Vision (ICCV'88)*, 1988. [Online]. Available: http://cim.mcgill.ca/clark/vmrl/web-content/papers/jjclark_iccv_1988.pdf

[38] P. a. Sandon, "Simulating visual attention," *Journal of Cognitive Neuroscience*, vol. 2, no. 3, p. 213–231, Jul. 1990. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/jocn.1990.2.3.213

[39] X. Gao, K. Hong, P. Passmore, L. Podladchikova, and D. Shaposhnikov, "Colour vision model-based approach for segmentation of traffic signs," *EURASIP Journal on Image and Video Processing*, 2008.

[40] D. Gao and N. Vasconcelos, "Integrated learning of saliency, complex features, and object detectors from cluttered scenes," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, vol. 2. IEEE, Jun. 2005, pp. 282–287 vol. 2.

[41] J. Tsotsos and A. Rothenstein, "Computational models of visual attention," *Scholarpedia*, vol. 6, no. 1, p. 6201, 2011. [Online]. Available: http://www.scholarpedia.org/article/Computational_models_of_visual_attention

[42] R. Rensink, J. O'Regan, and J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psychological Science*, 1997. [Online]. Available: http://pss.sagepub.com/content/8/5/368.short

[43] G. Buswell, *How people look at pictures: a study of the psychology and perception in art.* Oxford, England: Univ. Chicago Press, 1935. [Online]. Available: http://psych.wfu.edu/art_schirillo/articles/Buswell,1935.pdf

[44] C. Privitera and L. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, p. 970–982, 2000. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=877520

[45] J. Canny, "A computational approach to edge detection." *IEEE transactions on pattern analysis and machine intelligence*, vol. 8, no. 6, p. 679–98, Jun. 1986. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21869365

[46] E. Niebur and C. Koch, "Control of selective visual attention: Modeling the" where" pathway," *Advances in neural information processing systems*, p. 802–808, 1996.

[47] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time." *Vision research*, vol. 45, no. 5, p. 643–59, Mar. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15621181

[48] M. Pomplun, "Saccadic selectivity in complex visual search displays." *Vision research*, vol. 46, no. 12, p. 1886–900, Jun. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16445960

[49] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in neural information processing systems*, vol. 19, p. 545, 2007. [Online]. Available: http://papers.klab.caltech.edu/300/1/543.pdf

[50] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision research*, vol. 42, no. 1, p. 107–124, 2002. [Online]. Available: http://homepage.psy.utexas.edu/Homepage/Class/Psy394U/Hayhoe/perceptionaction/2008/week4readings/parkhurst.pdf

[51] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[52] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, Jun. 2007, pp. 1 –8.

[53] T. Kadir, D. Boukerroui, and M. Brady, "An analysis of the scale saliency algorithm," *OUEL No: 2264*, vol. 3, p. 1–38, 2003. [Online]. Available: http://www.hds.utc.fr/dboukerr/Papers/TR-2264-03.pdf

[54] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674 –693, Jul. 1989.

[55] Wikipedia, "Two-streams hypothesis," Feb. 2013, page Version ID: 540183557. [Online]. Available: http://en.wikipedia.org/w/index.php?title=Two-streams_hypothesis&oldid=540183557

[56] N. Ouerhani and H. Hugli, "Computing visual attention from scene depth," *Pattern Recognition, 2000. Proceedings. \ldots*, p. 375–378, 2000. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=905356

[57] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," *rn*, vol. 255, p. 3, 1999. [Online]. Available: https://svn.jde.gsyc.es/robotvision/trunk/papers/breazeal-context_dependent_attention-ijcai99.pdf

[58] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Handling uncertainty in video analysis with spatiotemporal visual attention," *The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ '05.*, p. 213–217, 2005. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1452395

[59] G. Heidemann, "Focus-of-attention from local color symmetries." *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 7, p. 817–30, Jul. 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18579942

[60] J. J. Koenderink and A. J. v. Doorn, "Representation of local geometry in the visual system," *Biological Cybernetics*, vol. 55, no. 6, pp. 367–375, Mar. 1987. [Online]. Available: http://link.springer.com/article/10.1007/BF00318371

[61] R. Linsker, "From basic network principles to neural architecture: emergence of orientation-selective cells." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 83, no. 21, p. 8390–4, Nov. 1986. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=386934&tool=pmcentrez&rendertype=abstract

[62] ——, "From basic network principles to neural architecture: emergence of spatial-opponent cells." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 83, no. 19, p. 7508–12, Oct. 1986. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=386934&tool=pmcentrez&rendertype=abstract

[63] ——, "From basic network principles to neural architecture: emergence of orientation columns." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 83, no. 22, p. 8779–83, Nov. 1986. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=387015&tool=pmcentrez&rendertype=abstract

[64] ——, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, p. 105–117, Mar. 1988. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=36

[65] Z. Li and L. Itti, "Gist based top-down templates for gaze prediction," in *Proc. Vision Science Society Annual Meeting (VSS09)*, May 2009, bu;td;mod;cv.

[66] T. A. Ell and S. J. Sangwine, "Hypercomplex fourier transforms of color images," *IEEE Transactions on Image Processing*, vol. 16, no. 1, p. 22–35, 2007. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4032812

[67] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Proc. European Conf. on Computer Vision*, 2012. [Online]. Available: http://cvhci.anthropomatik.kit.edu/~bschauer/pdf/schauerte2012quaternion-based.pdf

[68] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[69] A. Treisman, "Feature binding, attention and object perception," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 353, no. 1373, pp. 1295–1306, 1998.

[70] J. H. Maunsell and S. Treue, "Feature-based attention in visual cortex," *Trends in neurosciences*, vol. 29, no. 6, pp. 317–322, 2006.

[71] N. Bruce and P. Kornprobst, "On the role of context in probabilistic models of visual saliency," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, Nov. 2009, pp. 3089 –3092.

[72] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: a bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, 2008. [Online]. Available: http://w.journalofvision.org/content/8/7/32.short

[73] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802 –817, May 2006.

[74] F. Schumann, W. Einhäuser, J. Vockeroth, K. Bartl, E. Schneider, and P. König, "Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments," *Journal of Vision*, vol. 8, no. 14, Oct. 2008. [Online]. Available: http://www.journalofvision.org/content/8/14/12

[75] A. C. Le Ngo, G. Qiu, G. Underwood, L. M. Ang, and K. P. Seng, "Visual saliency based on fast nonparametric multidimensional entropy estimation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, p. 1305–1308. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6288129

[76] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," *Advances in neural information processing systems*, vol. 17, no. 481-488, p. 1, 2004. [Online]. Available: http://www.svcl.ucsd.edu/publications/conference/2004/nips04/nips04.pdf

[77] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychological Review*, vol. 113, no. 4, pp. 766–786, 2006.

[78] C. E. Shannon, "Prediction and entropy of printed english," *Bell system technical journal*, vol. 30, no. 1, pp. 50–64, 1951.

[79] ——, "The mathematical theory of communication," *M.D. computing : computers in medical practice*, vol. 14, no. 4, p. 306–17, 1948. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9230594

[80] G. Palm, "Evidence, information, and surprise," *Biological Cybernetics*, vol. 42, no. 1, pp. 57–68, Nov. 1981. [Online]. Available: http://www.springerlink.com/content/v8064m3454402894/

[81] J. Najemnik and W. S. Geisler, "Optimal eye movement strategies in visual search." *Nature*, vol. 434, no. 7031, p. 387–91, Mar. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15772663

[82] T. Lee and S. Yu, "An information-theoretic framework for understanding saccadic eye movements," *Advances in neural information processing systems*, 1999. [Online]. Available: http://www.cnbc.cmu.edu/tai/papers/nips99.pdf

[83] L. W. Renninger, J. Coughlan, P. Verghese, and J. Malik, "An information maximization model of eye movements." *Advances in neural information processing systems*, vol. 17, p. 1121–8, Jan. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16175670

[84] N. Brenner, W. Bialek, and R. de Ruyter van Steveninck, "Adaptive rescaling maximizes information transmission," *Neuron*, vol. 26, no. 3, pp. 695–702, Jun. 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0896627300812052

[85] W. E. Vinje and J. L. Gallant, "Natural stimulation of the nonclassical receptive field increases information transmission efficiency in v1," *The Journal of Neuroscience*, vol. 22, no. 7, pp. 2904–2915, Apr. 2002. [Online]. Available: http://www.jneurosci.org/content/22/7/2904

[86] G. Fritz, C. Seifert, L. Paletta, and H. Bischof, "Attentive object detection using an information theoretic saliency measure," *Attention and Performance in \ldots*, 2005. [Online]. Available: http://www.springerlink.com/index/NT451KWKFK4YU1XC.pdf

[87] T. Topper, "Selection mechanisms in human and machine vision," Ph.D. dissertation, University of Waterloo, 1991.

[88] D. Gao and N. Vasconcelos, "Discriminant interest points are stable," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007, p. 1–6. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4270146

[89] D. H. Hubel and T. N. WIESEL, "Receptive fields and functional architecture into nonstriate visual areas (18 and 19) of the cat." *Journal of neurophysiology*, vol. 28, p. 229–89, Mar. 1965. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14283058

[90] M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust, "Do we know what the early visual system does?" *The Journal of Neuroscience*, vol. 25, no. 46, pp. 10 577–10 597, Nov. 2005. [Online]. Available: http://www.jneurosci.org/content/25/46/10577

[91] W. S. Geisler and D. G. Albrecht, "Cortical neurons: isolation of contrast gain control," *Vision research*, vol. 32, no. 8, pp. 1409–1410, 1992.

[92] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of vision*, vol. 8, no. 7, 2008. [Online]. Available: http://ww.journalofvision.org/content/8/7/13.short

[93] P. Baldi and L. Itti, "Of bits and wows: A bayesian theory of surprise with applications to attention," *Neural Networks*, vol. 23, no. 5, p. 649–666, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608009003256

[94] K. Friston, "A theory of cortical responses," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1456, p. 815–836, 2005. [Online]. Available: http://rstb.royalsocietypublishing.org/content/360/1456/815.short

[95] K. Friston and S. Kiebel, "Predictive coding under the free-energy principle," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1521, pp. 1211–1221, May 2009, PMID: 19528002 PMCID: PMC2666703. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2666703/

[96] H. Feldman and K. J. Friston, "Attention, uncertainty, and Free-Energy," *Frontiers*

in Human Neuroscience, vol. 4, 2010. [Online]. Available: http://www.frontiersin.org/Human_Neuroscience/10.3389/fnhum.2010.00215/abstract

[97] M. I. Posner, *Chronometric explorations of mind*. Oxford, England: Lawrence Erlbaum, 1978, vol. xiii.

[98] C. Decaix, E. Siéroff, and P. Bartolomeo, "How voluntary is 'Voluntary' orienting of attention?" *Cortex*, vol. 38, no. 5, pp. 841–845, Jan. 2002. [Online]. Available: http://www.cortexjournal.net/article/S0010-9452(08)70053-4/abstract

[99] E. F. Risko and J. A. Stolz, "The proportion valid effect in covert orienting: strategic control or implicit learning?" *Consciousness and cognition*, vol. 19, no. 1, pp. 432–442, Mar. 2010, PMID: 20189414.

[100] K. Friston, "The free-energy principle: a unified brain theory?" *Nat Rev Neurosci*, vol. 11, no. 2, pp. 127–138, Feb. 2010. [Online]. Available: http://dx.doi.org/10.1038/nrn2787

[101] M. E. Mazurek, J. D. Roitman, J. Ditterich, and M. N. Shadlen, "A role for neural integrators in perceptual decision making," *Cerebral Cortex*, vol. 13, no. 11, pp. 1257–1269, Nov. 2003. [Online]. Available: http://cercor.oxfordjournals.org/content/13/11/1257

[102] R. C. Moeur, *Manual of traffic signs*. Online, 2003. [Online]. Available: http://www.ipl.org/IPL/Display?key=res:lii-16269&layout=blank

[103] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, ser. Pattern Classification and Scene Analysis: Pattern Classification, R. O. Duda, P. E. Hart, and D. G. Stork, Eds. Wiley, 2001, vol. 2, no. 6. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.133.1318&rep=rep1&type=pdf

[104] K. H. Lim, A. C. Le Ngo, K. P. Seng, and L. Ang, "UNMC-VIER AutoVision database," in *2010 International Conference on Computer Applications and Industrial Electronics (ICCAIE)*, Dec. 2010, pp. 650 –654. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5735015

[105] R. Kenneth, *Castleman, Digital image processing*. Prentice Hall Press, Upper Saddle River, NJ, 1996.

[106] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[107] K. H. Lim, K. P. Seng, A. C. Le Ngo, and L. M. Ang, "Real-time implementation of vision-based lane detection and tracking," in *Intelligent Human-Machine Systems and Cybernetics, 2009. IHMSC'09. International Conference on*, vol. 2, 2009, p. 364–367. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5335965

[108] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," *Advances in neural information processing systems*, vol. 20, p. 1–8, 2007. [Online]. Available: http://132.239.134.221/publications/conference/2007/nips2007/nips2007_budiscsal.pdf

[109] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Advances in neural information processing systems*, vol. 18, p. 547, 2006. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.69.9454&rep=rep1&type=pdf

[110] A. Cohen, I. Daubechies, and J. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 45, no. 5, p. 485–560, Jun. 1992. [Online]. Available: http://doi.wiley.com/10.1002/cpa.3160450502

[111] P. Viola and W. Wells III, "Alignment by maximization of mutual information," in *iccv*, vol. 24. Published by the IEEE Computer Society, 1995, p. 16. [Online]. Available: http://www.computer.org/portal/web/csdl/doi/10.1109/ICCV.1995.466930

[112] Z. Bylinskii, T. Judd, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark," http://saliency.mit.edu/.

[113] G. Qiu, X. Gu, Z. Chen, Q. Chen, and C. Wang, "An information theoretic model of spatiotemporal visual saliency," in *2007 IEEE International Conference on Multimedia and Expo*, Jul. 2007, pp. 1806 –1809.

[114] L. Sendur and I. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Transactions on Signal Processing*, vol. 50, no. 11, pp. 2744 – 2756, Nov. 2002.

[115] G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," *Computer Vision–ECCV 2008*, p. 44–57, 2008. [Online]. Available: http://www.springerlink.com/index/UR2062X5852X5355.pdf

[116] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," 2012. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6180177

[117] Y. Kim, A. Varshney, D. W. Jacobs, and F. Guimbretière, "Mesh saliency and human eye fixations," *ACM Transactions on Applied Perception*, vol. 7, no. 2, p. 1–13, Feb. 2010. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1670671.1670676

[118] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, p. 88–97, Jan. 2009. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167865508001220

[119] G. Kootstra, B. de Boer, and L. R. B. Schomaker, "Predicting eye fixations on complex visual stimuli using local symmetry," *Cognitive computation*, vol. 3, no. 1, p. 223–240, 2011. [Online]. Available: http://www.springerlink.com/index/36777502656UT607.pdf

[120] J. Starck and F. Murtagh, "Multiscale entropy filtering," *Signal Processing*, vol. 76, no. 2, pp. 147–165, Jul. 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168499000055

[121] P. Suau and F. Escolano, "A new feasible approach to multi-dimensional scale saliency," in *Advanced Concepts for Intelligent Vision Systems*, ser. Lecture Notes in Computer Science, J. Blanc-Talon, W. Philips, D. Popescu, and P. Scheunders, Eds. Springer Berlin Heidelberg, Jan. 2009, no. 5807, pp. 77–88. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-04697-1_8

[122] R. a. Andersen, R. M. Bracewell, S. Barash, J. W. Gnadt, and L. Fogassi, "Eye position effects on visual, memory, and saccade-related activity in areas LIP and 7a of macaque." *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 10, no. 4, p. 1176–96, Apr. 1990. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/2329374

[123] C. V. D. Malsburg, "The correlation theory of brain function," *Models of neural networks II*, no. July 1981, p. 1–26, 1994. [Online]. Available: http://books.google.com/books?hl=en&lr=&id=srw3tMi8HasC&oi=fnd&pg=PA95&dq=The+Correlation+Theory+of+Brain+Function&ots=YJQpgcPazQ&sig=IOIQK75bIjfrmgNNB_ee21n0U3ohttp://books.google.com/books?hl=en&lr=&id=srw3tMi8HasC&oi=fnd&pg=PA95&dq=The+correlation+theory+of+brain+function&ots=YJQpgcPaAR&sig=W-7gQoCaufw6lXvTWfZq4hFVaXw

[124] A. B. Watson, "The cortex transform: Rapid computation of simulated neural images," *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 311–327, Sep. 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0734189X87801846

[125] H. Senane, A. Saadane, and D. Barba, "Image coding in the context of a psycho-visual image representation with vector quantization," in , *International Conference on Image Processing, 1995. Proceedings*, vol. 1, Oct. 1995, pp. 97 –100 vol.1.

[126] M. Unser, N. Chenouard, and D. Van De Ville, "Steerable pyramids and tight wavelet frames in L_2(BBR^d)," *IEEE Transactions on Image Processing*, vol. 20, no. 10, pp. 2705 –2721, Oct. 2011.

[127] M. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2091 –2106, Dec. 2005.

[128] W. L. Chan, H. Choi, and R. G. Baraniuk, "Coherent multiscale image processing using dual-tree quaternion wavelets," *Image Processing, IEEE Transactions on*, vol. 17, no. 7, p. 1069–1082, 2008. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4526699

[129] S. Mallat, *A wavelet tour of signal processing*. San Diego, CA: Academic press, 1999. [Online]. Available: http://books.google.com.my/books?hl=en&lr=&id=hbVOfWQNtB8C&oi=fnd&pg=PP2&dq=A+wavelet+tour+of+signal+processing&ots=qrqZqpjY0-&sig=gW58pBKHeA-D2CkhBMgYCPUfVHI

[130] R. Coifman and M. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713 –718, Mar. 1992.

[131] T. Bulow and G. Sommer, "Hypercomplex signals-a novel extension of the analytic signal to the multidimensional case," *IEEE Transactions on Signal Processing*, vol. 49, no. 11, pp. 2844 –2852, Nov. 2001.

[132] N. G. Kingsbury, "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters," in *Proc. 8th IEEE DSP Workshop*, vol. 8, 1998, p. 86. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.51.1213&rep=rep1&type=pdf

[133] N. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Applied and Computational Harmonic Analysis*, vol. 10, no. 3, p. 234–253,

2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1063520300903439

[134] I. Bayram and I. W. Selesnick, "On the dual-tree complex wavelet packet and m-band transforms," *IEEE transactions on signal processing*, vol. 56, no. 6, p. 2298–2310, 2008. [Online]. Available: http://cat.inist.fr/?aModele=afficheN&cpsidt=20350138

[135] I. Cohen, S. Raz, and D. Malah, "Orthonormal shift-invariant wavelet packet decomposition and representation," *Signal Processing*, vol. 57, no. 3, pp. 251–270, Mar. 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168497000078

[136] T. Kadir, D. Boukerroui, and M. Brady, "An analysis of the scale saliency algorithm," *OUEL No: 2264*, vol. 3, 2003. [Online]. Available: http://www.robots.ox.ac.uk/~timork/Saliency/TR-2264-03.pdf

[137] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform," *Signal Processing Magazine, IEEE*, vol. 22, no. 6, p. 123–151, 2005. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1550194

[138] A. Borji, D. N. Sihite, and L. Itti, "An Object-Based bayesian framework for Top-Down visual attention," in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/download/5016/5297

[139] M. Crouse and R. Baraniuk, "Simplified wavelet-domain hidden markov models using contexts," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998*, vol. 4, May 1998, pp. 2277 –2280 vol.4.

[140] J. K. Romberg, H. Choi, R. G. Baraniuk, and N. G. Kingsbury, "Hidden markov tree models for complex wavelet transforms," *IEEE Transactions on Signal Processing*, 2002. [Online]. Available: http://dspaceland.rice.edu/bitstream/handle/1911/4447/Rom2002May1HiddenMark.PDF?sequence=1

[141] H. Choi and R. Baraniuk, "Multiscale texture segmentation using wavelet-domain hidden markov models," in *Conference Record of the Thirty-Second Asilomar Conference on Signals, Systems amp; Computers, 1998*, vol. 2, Nov. 1998, pp. 1692 –1697 vol.2.

[142] D. Zoran and Y. Weiss, "Scale invariance and noise in natural images," in *2009 IEEE 12th International Conference on Computer Vision*, Oct. 2009, pp. 2209

–2216. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5459476

[143] Y. Weiss and W. Freeman, "What makes a good model of natural images?" in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, Jun. 2007, pp. 1 –8.

[144] G. A. Alvarez and A. Oliva, "Spatial ensemble statistics are efficient codes that can be represented with reduced attention," *Proceedings of the National Academy of Sciences*, vol. 106, no. 18, pp. 7345–7350, May 2009. [Online]. Available: http://www.pnas.org/content/106/18/7345

[145] M. N. Do and M. Vetterli, "Rotation invariant texture characterization and retrieval using steerable wavelet-domain hidden markov models," *Multimedia, IEEE Transactions on*, vol. 4, no. 4, p. 517–527, 2002. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1176950

[146] ——, "Wavelet-based texture retrieval using generalized gaussian density and Kullback-Leibler distance." *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 11, no. 2, p. 146–58, Jan. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18244620

[147] N. Minh, "Directional multiresolution image representations," Ph.D. dissertation, Citeseer, 2002. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.162.3137&rep=rep1&type=pdf

[148] K. Shi and S. Zhu, "Mapping natural image patches by explicit and implicit manifolds," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, Jun. 2007, pp. 1 –7.

[149] J. Romberg, H. Choi, and R. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden markov models," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 1056 –1068, Jul. 2001.

[150] H. Choi and R. Baraniuk, "Multiscale image segmentation using wavelet-domain hidden markov models," *IEEE Transactions on Image Processing*, vol. 10, no. 9, pp. 1309 –1321, Sep. 2001.

[151] D. Y. P. Duncan and M. N. Do, "Directional multiscale statistical modeling of images," *Urbana*, vol. 51, p. 61801, 2003. [Online]. Available: http://spiedigitallibrary.org/data/Conferences/SPIEP/20644/69_1.pdf

[152] E. Simoncelli and J. Portilla, "Texture characterization via joint statistics of wavelet coefficient magnitudes," in *1998 International Conference on Image Processing, 1998. ICIP 98. Proceedings*, vol. 1, Oct. 1998, pp. 62 –66 vol.1.

[153] H. Cheng and C. Bouman, "Trainable context model for multiscale segmentation," in *1998 International Conference on Image Processing, 1998. ICIP 98. Proceedings*, vol. 1, Oct. 1998, pp. 610 –614 vol.1.

[154] J. Li, R. Gray, and R. Olshen, "Multiresolution image classification by hierarchical modeling with two-dimensional hidden markov models," *IEEE Transactions on Information Theory*, vol. 46, no. 5, pp. 1826 –1841, Aug. 2000.

[155] C. Bouman and M. Shapiro, "A multiscale random field model for bayesian image segmentation," *IEEE Transactions on Image Processing*, vol. 3, no. 2, pp. 162 –177, Mar. 1994.

[156] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a bayesian approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 60, no. 4, p. 725–749, 1998. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00151/abstract

[157] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 909 –919, Apr. 1999.

[158] H. A. Chipman, E. D. Kolaczyk, and R. E. McCulloch, "Adaptive bayesian wavelet shrinkage," *Journal of the American Statistical Association*, vol. 92, no. 440, pp. 1413–1421, 1997. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1997.10473662

[159] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of Human-Model agreement in visual saliency modeling: A comparative study," *Image Processing, IEEE Transactions on*, 2012. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6253254

[160] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs," *Current Opinion in Neurobiology*, vol. 14, no. 4, p. 481–7, Aug. 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15321069