

# **Learning Feature Selection and Combination Strategies for Generic Salient Object Detection**

by

Syed Saud Naqvi

A thesis  
submitted to the Victoria University of Wellington  
in fulfilment of the  
requirements for the degree of  
Doctor of Philosophy  
in Computer Science.

Victoria University of Wellington  
2016



## Abstract

For a diverse range of applications in machine vision from social media searches to robotic home care providers, it is important to replicate the mechanism by which the human brain selects the most important visual information, while suppressing the remaining non-usable information.

Many computational methods attempt to model this process by following the *traditional model of visual attention*. The traditional model of attention involves feature extraction, conditioning and combination to capture this behaviour of human visual attention. Consequently, the model has inherent design choices at its various stages. These choices include selection of parameters related to the feature computation process, setting a conditioning approach, feature importance and setting a combination approach. Despite rapid research and substantial improvements in benchmark performance, the performance of many models depends upon tuning these design choices in an *ad hoc* fashion. Additionally, these design choices are heuristic in nature, thus resulting in good performance only in certain settings. Consequentially, many such models exhibit low robustness to difficult stimuli and the complexities of real-world imagery.

Machine learning and optimisation technique have long been used to increase the generalisability of a system to unseen data. Surprisingly, artificial learning techniques have not been investigated to their full potential to improve generalisation of visual attention methods.

The proposed thesis is that artificial learning can increase the generalisability of the traditional model of visual attention by effective selection and optimal combination of features.

The following new techniques have been introduced at various stages

of the traditional model of visual attention to improve its generalisation performance, specifically on challenging cases of saliency detection:

1. *Joint optimisation of feature related parameters and feature importance weights is introduced for the first time to improve the generalisation of the traditional model of visual attention.* To evaluate the joint learning hypothesis, a new method namely GAOVSM is introduced for the tasks of eye fixation prediction. By finding the relationships between feature related parameters and feature importance, the developed method improves the generalisation performance of baseline method (that employ human encoded parameters).
2. Spectral matting based figure-ground segregation is introduced to overcome the artifacts encountered by region-based salient object detection approaches. By suppressing the unwanted background information and assigning saliency to object parts in a uniform manner, the developed FGS approach overcomes the limitations of region based approaches.
3. *Joint optimisation of feature computation parameters and feature importance weights is introduced for optimal combination of FGS with complementary features for the first time for salient object detection.* By learning feature related parameters and their respective importance at multiple segmentation thresholds and by considering the performance gaps amongst features, the developed FGSopt method improves the object detection performance of the FGS technique also improving upon several state-of-the-art salient object detection models.
4. The introduction of multiple combination schemes/rules further extends the generalisability of the traditional attention model beyond that of joint optimisation based single rules. The introduction of feature composition based grouping of images, enables the developed IGA method to autonomously identify an appropriate combination



strategy for an unseen image. The results of a pair-wise ranksum test confirm that the IGA method is significantly better than the deterministic and classification based benchmark methods on the 99% confidence interval level. Extending this line of research, a novel relative encoding approach enables the adapted XCSCA method to group images having similar saliency prediction ability. By keeping track of previous inputs, the introduced action part of the XCSCA approach enables learning of generalised feature importance rules. By more accurate grouping of images as compared with IGA, generalised learnt rules and appropriate application of feature importance rules, the XCSCA approach improves upon the generalisation performance of the IGA method.

5. The introduced uniform saliency assignment and segmentation quality cues enable label free evaluation of a feature/saliency map. By accurate ranking and effective clustering, the developed DFS method successfully solves the complex problem of finding appropriate features for combination (on an-image-by-image basis) for the first time in saliency detection. The DFS method enables ground truth free evaluation of saliency methods and advances the state-of-the-art in data driven saliency aggregation by detection and deselection of redundant information.

The final contribution is that the developed methods are formed into a complete system where analysis shows the effects of their interactions on the system. Based on the saliency prediction accuracy versus computational time trade-off, specialised variants of the proposed methods are presented along with the recommendations for further use by other saliency detection systems.

This research work has shown that artificial learning can increase the generalisation of the traditional model of attention by effective selection and optimal combination of features. Overall, this thesis has shown that it

is the ability to autonomously segregate images based on their types and subsequent learning of appropriate combinations that aid generalisation on difficult unseen stimuli.

# List of Publications

Parts of research contributions have been published/submitted in the following conferences and journals:

1. Syed S. Naqvi, Will N. Browne, and Christopher Hollitt, "Feature Quality Based Dynamic Feature Selection for Improving Salient Object Detection", *IEEE Transactions on Image Processing*, 2015, *Provisionally Accepted*.
2. Syed S. Naqvi, Will N. Browne, and Christopher Hollitt, "Salient Object Detection Via Spectral Matting", *Pattern Recognition*, 2015, *Conditionally Accepted*.
3. Syed S. Naqvi, Will N. Browne, and Christopher Hollitt, "Evolutionary Feature Combination Based Seed Learning for Diffusion-Based Saliency", in *Proceedings of the Simulated Evolution and Learning*, 2014, 822-834.
4. Muhammad Iqbal, Syed Saud Naqvi, Will N. Browne, Christopher Hollitt and Mengjie Zhang, "Salient Object Detection Using Learning Classifier Systems That Compute Action Mappings", in *Proceedings of the Conference on Genetic and Evolutionary Computation*, 2014, 525-532. This work was awarded best paper award in the Evolutionary Machine Learning track.
5. Syed S. Naqvi, Will N. Browne, and Christopher Hollitt, "Genetic Algorithms Based Feature Combination For Salient Object Detection,

For Autonomously Identified Image Domain Types”, in *Proceedings of the IEEE Congress on Evolutionary Computation*, 2014, 109-116.

6. Syed S. Naqvi, Will N. Browne, and Christopher Hollitt, “Combining Object-Based Local and Global Feature Statistics For Salient Object Search”, in *Proceedings of the International Conference of Image and Vision Computing*, 2013, 394-399.
7. Syed S. Naqvi, Will N. Browne, and Christopher Hollitt, “Optimizing Visual Attention Models For Predicting Human Fixations Using Genetic Algorithms”, in *Proceedings of the IEEE Congress on Evolutionary Computation*, 2013, 1302-1309.
8. Syed S. Naqvi, Will N. Browne, and Christopher Hollitt, “Optimizing Bio-Inspired Visual Attention Model Using Genetic Algorithm For Predicting Human Fixations”, in *NZCSRSC*, 2013. This work was awarded best paper award and a prize money of 1500 NZD as a travel grant.

# Acknowledgments

All praise unto GOD, Who made it easier for me to conceive this work. After that, I am thankful to my supervisors, Dr. Will Browne and Dr. Christopher Hollitt, for helping me with my research, presentation and personal skills. I feel very lucky to have supervisors who spent long dedicated hours to review my writings even on weekends. Their corrections and suggestions have greatly helped me in improving my writing skills over the past three years. I am grateful to them for all the help with structuring and writing of my thesis and last but not the least, the numerous proof reading iterations.

I would like to thank Will Browne for all the advice on academic and non academic matters that I sought on many occasions; sometimes even on tiny matters. I appreciate the way he related complex problems to the real world to make me understand better and for suggesting practical solutions to research problems. I would also like to thank Will for helping me with the numerous practice talks and providing reference and progress letters on multiple occasions. Thank you Will for helping me with the tutoring contract and the URF to make my stay in New Zealand possible until my oral defence. I would like to thank Christopher Hollitt for his advice on all matters. I acknowledge his help to improve my understanding of the human visual system with reference to active vision processes. Thank you Chris for helping me with the notation in several papers and the thesis. The regular meetings in a friendly environment at the ECRG research group helped me in progressing well in my studies and also contributed

to my understanding of various aspects of evolutionary computation.

I am especially thankful to Victoria University of Wellington for awarding me the Victoria Doctoral Scholarship and to both my supervisors for supporting my application. I am also highly thankful to the Hardship Fund for helping me on multiple occasions.

I acknowledge the help and support of my parents and family members back home. I also acknowledge the help and support of my wife, who took care of our children while I was busy in my studies and was not able to give them their deserved time.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Scope . . . . .	1
1.2	Motivation . . . . .	3
1.2.1	Why Genetic Algorithm? . . . . .	8
1.2.2	Why Learning Classifier Systems? . . . . .	9
1.3	Thesis Statement . . . . .	10
1.4	Thesis Goals . . . . .	10
1.5	Thesis Contributions . . . . .	13
1.6	Thesis Organisation . . . . .	15
<b>2</b>	<b>Background</b>	<b>19</b>
2.1	Computational Methods of Visual Attention . . . . .	19
2.1.1	General Structure of Computational Models . . . . .	20
2.1.2	Choice of Parameters to be Optimised . . . . .	29
2.2	Salient Object Detection . . . . .	30
2.2.1	Methods for Salient Object Detection . . . . .	32
2.3	Machine Learning Methods . . . . .	43
2.3.1	Genetic Algorithms . . . . .	44
2.3.2	Learning Classifier Systems . . . . .	46
2.4	Chapter Summary . . . . .	49
<b>3</b>	<b>Learning Visual Attention</b>	<b>51</b>
3.1	Introduction . . . . .	51

3.1.1	Chapter Organisation . . . . .	54
3.2	GAOVSM . . . . .	55
3.2.1	The Visual Saliency Method . . . . .	55
3.2.2	Genetic Algorithm . . . . .	56
3.3	Design of Experiments . . . . .	60
3.3.1	Data Set . . . . .	60
3.3.2	Ground Truth Data . . . . .	61
3.3.3	Performance Measures . . . . .	61
3.3.4	Experimental Settings for the Genetic Algorithm . . .	62
3.3.5	Selected Visual Attention Methods for Comparison .	63
3.4	Results and Discussion . . . . .	63
3.4.1	Comparison of learned versus baseline features . . .	64
3.4.2	Analysis of Evolved Solutions . . . . .	64
3.4.3	Comparison of GAOVSM with State-of-the-art . . . .	68
3.5	Chapter Summary . . . . .	72
<b>4</b>	<b>Spectral Matting Based Object Detection</b>	<b>75</b>
4.1	Introduction . . . . .	75
4.1.1	Chapter Goals . . . . .	79
4.1.2	Chapter Organisation . . . . .	80
4.2	Figure-ground Segregation Method . . . . .	80
4.2.1	Colour Spatial Distribution . . . . .	81
4.2.2	Eigenvectors of the Matting Laplacian . . . . .	83
4.2.3	Matting Components from Eigenvectors . . . . .	83
4.2.4	Foreground Object Saliency . . . . .	88
4.3	Design of Experiments . . . . .	93
4.3.1	Dataset . . . . .	93
4.3.2	Parameter Settings . . . . .	93
4.3.3	Evaluation Benchmarks . . . . .	94
4.4	Results and Discussion . . . . .	94
4.4.1	Discussion of FGS results . . . . .	95



4.5	Chapter Summary . . . . .	98
<b>5</b>	<b>Improving FGS</b>	<b>101</b>
5.1	Introduction . . . . .	101
5.1.1	Chapter Goals . . . . .	103
5.1.2	Chapter Organisation . . . . .	104
5.2	Method . . . . .	104
5.2.1	Optimisation Framework . . . . .	106
5.2.2	Final Saliency Computation . . . . .	111
5.3	Design of Experiments . . . . .	111
5.3.1	Datasets . . . . .	111
5.3.2	Parameter Settings . . . . .	113
5.3.3	Evaluation Benchmarks . . . . .	113
5.4	Results and Discussion . . . . .	114
5.4.1	Comparison with the FGS method . . . . .	114
5.4.2	Comparison with the Baseline Unoptimised method . . . . .	114
5.4.3	Comparison with the State-of-the-art methods . . . . .	118
5.4.4	Interpretation of Results . . . . .	121
5.5	Chapter Summary . . . . .	128
<b>6</b>	<b>Genetic Algorithm For Feature Combination</b>	<b>131</b>
6.1	Introduction . . . . .	131
6.2	Rationale for Learning Multiple Schemes . . . . .	133
6.3	Proposed Methods . . . . .	136
6.3.1	Feature Extraction . . . . .	137
6.3.2	Genetic Algorithm . . . . .	140
6.3.3	Image Dependent GA Based Approach . . . . .	143
6.4	Design of Experiments . . . . .	146
6.4.1	Datasets and Experimental Setup . . . . .	146
6.4.2	Selected methods for Comparison . . . . .	148
6.4.3	Training Experiments . . . . .	149
6.5	Results and Discussion . . . . .	150

6.5.1	Comparison of IGA with the Baseline GA . . . . .	150
6.5.2	Comparison of IGA with Existing Work . . . . .	152
6.5.3	Qualitative Comparison . . . . .	154
6.5.4	Analysis of Evolved Solutions . . . . .	154
6.6	Chapter Summary . . . . .	160
<b>7</b>	<b>Learning Classifier Systems Based Combination</b>	<b>163</b>
7.1	Introduction . . . . .	163
7.2	Salient Object Detection Using Learning Classifier Systems .	164
7.3	Experimental Design . . . . .	167
7.3.1	Data Set . . . . .	167
7.3.2	Experimental Setup . . . . .	169
7.4	Results and Discussions . . . . .	170
7.5	Analysis of Evolved Solutions . . . . .	172
7.6	Analysis of Grouping Schemes . . . . .	175
7.7	Further Discussions . . . . .	178
7.8	Chapter Summary . . . . .	180
<b>8</b>	<b>Dynamic Feature Selection</b>	<b>183</b>
8.1	Introduction . . . . .	183
8.1.1	Chapter Goals . . . . .	184
8.1.2	Chapter Organisation . . . . .	185
8.2	Graph Based Manifold Ranking . . . . .	186
8.3	Dynamic Feature Selection (DFS) method . . . . .	186
8.3.1	System Model . . . . .	187
8.3.2	Foreground Approximation . . . . .	187
8.3.3	Cues for Measuring Feature Quality . . . . .	191
8.3.4	Objective Function for Feature Quality . . . . .	198
8.3.5	Feature Selection Algorithm . . . . .	200
8.3.6	Feature/Saliency Fusion . . . . .	203
8.4	Experimental Setup . . . . .	204
8.5	Results and Discussion . . . . .	207

8.5.1	Feature Selection for Feature Integration . . . . .	209
8.5.2	Saliency Aggregation . . . . .	212
8.6	Results of the DFS Based Aggregation . . . . .	214
8.6.1	Further Discussion . . . . .	215
8.7	Ranking Evaluation . . . . .	217
8.8	Chapter Summary . . . . .	220
<b>9</b>	<b>Discussions</b>	<b>223</b>
9.1	DFS in Learned Combination . . . . .	225
9.2	FGS as Foreground Approximation . . . . .	230
9.3	Improving State-of-the-art Methods . . . . .	234
9.3.1	FGS improving benchmark methods . . . . .	235
9.3.2	DFS Improving Benchmark Methods . . . . .	241
9.4	Recommendations for Salient Object Detection Methods . .	246
9.4.1	Incorporating Dynamic feature Selection . . . . .	246
9.4.2	Incorporating Learning Methods . . . . .	249
<b>10</b>	<b>Conclusions and Future Work</b>	<b>255</b>
10.1	Achieved Objectives . . . . .	255
10.2	Conclusions . . . . .	258
10.2.1	Joint Optimisation . . . . .	258
10.2.2	Multiple Combination Schemes . . . . .	261
10.2.3	Dynamic Feature Selection (DFS) . . . . .	263
10.3	Future Work . . . . .	264
10.3.1	Matting Based Figure-ground Segregation . . . . .	264
10.3.2	Learning Multiple Combination Rules . . . . .	264
10.3.3	Dynamic Feature Selection . . . . .	265



# List of Figures

1.1	The traditional computational model of visual attention . . .	4
1.2	Structure of the contributions chapters . . . . .	16
2.1	Computational model of visual attention . . . . .	21
2.2	Effects of the wavelength ( $\lambda$ ) and elongation ( $\gamma$ ) . . . . .	23
2.3	Orientations and phase of the Gabor filter . . . . .	24
2.4	Importance of the normalisation scheme w.r.t input type . .	27
2.5	Choice of integration operation . . . . .	28
2.6	Difference between salient object detection and figure-ground segmentation . . . . .	31
2.7	Artifacts of region based processing . . . . .	35
2.8	Shortcomings of feature selection approaches . . . . .	40
3.1	Computational model of visual attention . . . . .	52
3.2	Comparison of learned versus baseline features . . . . .	65
3.3	Effect of parameters on the orientation feature . . . . .	67
3.4	Performance comparison of methods . . . . .	70
3.5	Qualitative comparison of visual methods . . . . .	73
4.1	Artifacts of region based approaches . . . . .	77
4.2	System model of the FGS . . . . .	81
4.3	Convergence plot of the mean sparsity score . . . . .	86
4.4	Effects of image sizes on performance . . . . .	87
4.5	Comparison of FGS with region based approaches . . . . .	95

4.6	Visual comparison of FGS with state-of-the-art . . . . .	96
4.7	Object coverage of matting components . . . . .	97
4.8	Foreground matting components selection . . . . .	98
4.9	Rationale of learning feature importance . . . . .	99
5.1	System model of FGSopt . . . . .	105
5.2	Rationale for learning feature importance . . . . .	107
5.3	Rationale for learning feature related parameters . . . . .	109
5.4	Comparison of feature level performance with FGSopt . . . . .	115
5.5	Comparison of FGSopt with unoptimised method . . . . .	116
5.6	PR curves comparison with benchmark methods . . . . .	120
5.7	Precision, recall and F-measure comparison . . . . .	122
5.8	FGSopt induced segmentations . . . . .	123
5.9	Visual comparison with the state-of-the-art approaches . . . . .	126
6.1	Rationale for grouping images based on image types . . . . .	135
6.2	System model of IGA . . . . .	137
6.3	A pattern of the phenotype of IGA . . . . .	141
6.4	Effect of nearest neighbours $k$ on grouping performance . . . . .	149
6.5	Quantitative comparison of baseline GA with IGA method . . . . .	151
6.6	Quantitative comparison of IGA with learning benchmarks . . . . .	153
6.7	Visual comparison of IGA with benchmark methods . . . . .	155
6.8	Salient object segmentation induced by saliency methods . . . . .	156
6.9	Explanation of multiple learned weights by IGA . . . . .	158
6.10	Effects of learned normalisation scheme . . . . .	159
7.1	Encoding scheme to match input with classifier population . . . . .	166
7.2	Training process of the XCSCA method . . . . .	170
7.3	Comparison of the XCSCA with IGA method . . . . .	171
7.4	Visual comparison of XCSCA with IGA . . . . .	171
7.5	Comparison of grouping performance of XCSCA and IGA . . . . .	176
7.6	Explanation of the intergroup variance . . . . .	178

7.7	Categorical results of the IGA and XCSCA . . . . .	179
8.1	Comparison of image features using XCS's encoding . . . . .	184
8.2	System model for dynamic feature selection . . . . .	188
8.3	Rationale for uniform saliency cues . . . . .	193
8.4	Comparison of $e_{\text{contour}}$ measure with $f_B$ . . . . .	197
8.5	Distributions of the proposed quality measurement . . . . .	199
8.6	Example of clustering for feature selection . . . . .	202
8.7	Parametric effects on $\mathbf{F}$ . . . . .	208
8.8	Quantitative results for multi-level saliency fusion . . . . .	211
8.9	Visual comparison of multi-level saliency fusion . . . . .	212
8.10	Quantitative results for saliency aggregation . . . . .	213
8.11	Visual comparison of DFS with PW . . . . .	214
8.12	Comparison of DFS with benchmark methods . . . . .	216
8.13	Visual comparison of DFS with benchmarks . . . . .	217
8.14	Performance comparison of DFS on ranking . . . . .	219
8.15	Representative visual example of ranking results . . . . .	219
9.1	Overview of the thesis structure . . . . .	224
9.2	Effect of DFS on combination approaches . . . . .	226
9.3	Visualisation of dynamic feature selection . . . . .	228
9.4	Illustrations of clustering based on feature quality . . . . .	229
9.5	Visual comparison of $(\mathbf{F}_G)$ with FGS . . . . .	232
9.6	AUCPR comparison of foreground approximation methods . . . . .	233
9.7	Timing comparison of foreground approximation methods . . . . .	233
9.8	FGS improving benchmark methods . . . . .	236
9.9	Visual comparison of FGS with FGS improved methods . . . . .	239
9.10	Performance improvements by FGS on two images . . . . .	240
9.11	Precision-recall curves for experiment I . . . . .	243
9.12	Visual comparison of DFS with GC and SISal . . . . .	245
9.13	Guidelines for salient object detection methods . . . . .	247





# List of Tables

3.1	Optimised parameters $\phi$ evolved by GA . . . . .	66
5.1	Feature level improvements by optimised parameters . . . .	117
5.2	Timing comparison with learning based benchmark methods	127
6.1	Statistical comparisons of IGA with other methods . . . . .	153
6.2	Representative learned solutions by the IGA method . . . .	155
6.3	Representative learned solutions with GA method . . . . .	157
7.1	Representative solutions learned by IGA . . . . .	173
7.2	Sample of the experienced and accurate classifiers . . . . .	174



# Chapter 1

## Introduction

Humans are exquisitely efficient in selecting the most important visual information from a scene and suppressing the remaining non-usable information. The means by which the brain prioritises the incoming information is called “*selective attention*”. Replicating this ability in machines would foster a diverse range of applications from social media searches to robotic home care providers. However current machine vision techniques do not exploit this ability to its full potential and need improvement.

### 1.1 Scope

Machine vision encompasses the design and implementation of systems that allow machines to recognise objects and transform the incoming information, such that it is adequate for tasks typically requiring human vision. It includes detecting salient objects in a scene/image.

Most of the applications in machine vision have to deal with large amounts of data, so the computational complexity related to image interpretation in such applications is high [131]. Despite active research in machine vision and robotics, many real-world tasks such as object of interest detection and search, which are easily performed by humans, are still challenging for machines. In order to deal with these demands, researchers

in computer vision, machine vision and graphics have investigated the exploitation of concepts from human selective attention to prioritise information. This reduces the processing of irrelevant data and hence improves the computational load of subsequent algorithms. The usage of artificial systems that attempt to mimic human selective attention has led to the development of computational models of visual attention.

The majority of current methods introduced for modelling visual attention follow the traditional model of visual attention [66]. These methods compute multiple features at the first stage of processing and subsequently integrate them to obtain a so-called saliency map that highlights objects or regions in the input image. Despite intense research and substantial improvements in benchmark performance, the performance of many models depends on tuning of parameters in an *ad hoc* fashion. Additionally, the design choices are heuristic in nature, often resulting in good performance in only certain settings. Occasionally, the poor performance of heuristic methods may be attributed to the expectation that a heuristic will fit to all situations. Consequentially, many such models exhibit low robustness to difficult stimuli and struggle to generalise to challenging cases of saliency detection, such as similar foreground and background, cluttered background and multiple salient objects.

Machine learning and optimisation technique have long been used to increase the generalisability of a system to unseen data [142]. Surprisingly, artificial learning techniques have not been investigated to their full potential to improve generalisation of visual attention methods. Therefore, new methods are needed to incorporate artificial learning at various stages of the visual attention model in order to extend its generalisability to challenging cases in machine vision.

## 1.2 Motivation

There are a number of factors that affect the performance of a computational method for visual attention. On a broad scale these factors include, important parameters of the feature computation process, the functions chosen to condition features, assignment of importance to features and the function chosen to combine features, have a great influence on the overall performance of a saliency model. A general structure of the traditional computational model of visual attention that was introduced by Itti et al. [66] is shown in Figure 1.1. It has three stages as labelled in the figure. The first stage involves feature extraction (along with choice of feature related parameters), the second stage performs feature conditioning using a normalisation function and the final stage is responsible for weighted combination of the feature maps to construct the saliency map. Many methods that follow the traditional model of Figure 1.1, employ human encoded parameters and region based feature computation at stage 1, a fixed normalisation function at stage 2, neglect relative feature importance and adopt a heuristic combination function at stage 3. These fixed and heuristic choices often lead to poor generalisability in current methods.

The traditional computational model of attention requires numerous design choices, such as the number of layers in the image pyramid, types of image pyramids, number and parameters of orientation filters, designation of centre and surround levels, feature importance weights, etc. These parameters greatly affect the performance of the overall computational method for visual attention. One important question is therefore how to learn the best suited options for these various design choices, as this has been not been systematically investigated in previous work.

For the first two stages of processing, the traditional computational model of visual attention employed a  $5 \times 5$  filter size to compute the image pyramid (Figure 1.1). In contrast, Frintrop [43] found a  $3 \times 3$  convolution

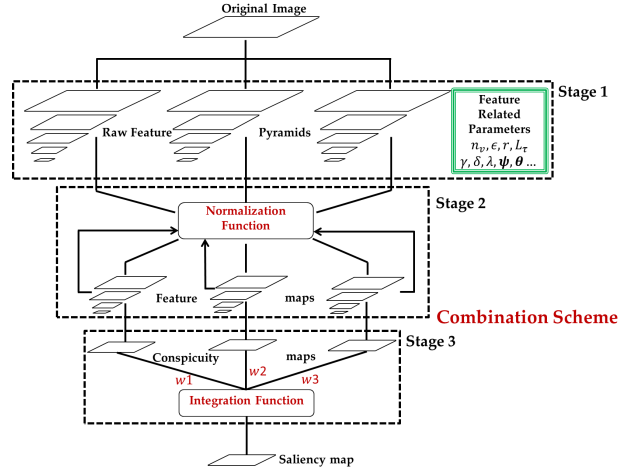


Figure 1.1: The traditional computational model of visual attention proposed by Itti et al. [66]. The factors that influence performance are highlighted at each stage. The normalisation function in stage 2, weights ( $w_1, w_2, w_3$ ) and the integration function in stage 3 constitute a combination scheme.

filter to be more useful in her experiments, while in another study Frin-trop et al. [44] found a filter size of  $15 \times 15$  pixels to work well. Itti et al. proposed the use of nine spatial scales for the image pyramid, while Klein and Frin-trop [77] found 8 spatial scales to work well in their experiments. For the orientation channel, the traditional computational model of attention [66] employed four orientations for the filters ( $0^\circ, 45^\circ, 90^\circ, 135^\circ$ ). Wischnewski et al. [141] found five orientations ( $0^\circ, 36^\circ, 72^\circ, 108^\circ, 144^\circ$ ) to be more useful, while Bian and Zhang [112] proposed unique orientations, i.e.  $0^\circ, 4^\circ, 8^\circ, 16^\circ$ .

At the final stage of the saliency computation process (see Figure 1.1), Zhao et al. [153] found the orientation feature to be more informative than intensity and colour features for eye fixation prediction. Judd [71] reported colour to be a better predictor of human eye fixations than several other features, while Cerf et al. [23] found that people looked more at faces (in their experiments), making the face channel more important than inten-

sity, colour and orientation features.

The large number of design choices and complex interactions between various parameters of the model cause the search space to grow enormously. In most cases, it becomes impractical to exhaustively sweep through the entire search space and an efficient search approach appears to be necessary to address this problem. A few methods [65, 72] attempt to learn the feature importance to improve generalisation. However these methods use human encoded parameters for the majority of the design choices. As these methods employ classification techniques to learn feature importance, it is not feasible for them to perform parameter search. By only learning feature importance, they neglect the importance of parameter tuning and fail to capture the interactions between the different parameters that may help in improving generalisation. Genetic Algorithms are global optimisation methods that can jointly search the important parameters and the feature importance such that the difference between the model's prediction and the target output is minimised (please see section 1.2.1 for further details.). The optimised parameters can then be employed by the model to increase generalisation on unseen images.

An important related question is, which features can best enable the system to generalise on challenging cases of saliency detection. A vast variety of bottom-up and top-down features have been investigated for eye fixation prediction and found to be very useful. In the case of salient object detection, comparatively fewer features are proposed and region based approaches strongly dominate. Region based approaches refer to those techniques that divide the image by grouping its pixels into small homogeneous regions before saliency computation. The shortcomings of these models are that they fail to uniformly highlight the salient (target) object and can completely miss parts of the salient objects in the final saliency map. Dividing the images into regions before saliency computation is the cause of these undesired artifacts where object boundaries are not respected and objects are cut into pieces. The recent approach of Ming et al.

[26], termed as SIA, attempted to overcome the problem of non-uniform saliency assignment by replacing hard object boundaries with soft image abstractions. The approach of Ming et al. groups initial hard clusters into more semantically meaningful clusters without specifying the number of clusters. The final clusters are formed based on the underlying image distribution. However, the response of SIA is still subject to non-uniform saliency assignment when the regions resulting from soft abstractions also fail to respect object boundaries.

A potential solution to the non-uniform saliency inside objects can be inspired by image matting. Matting is the process of blocking out unwanted information in the visual scene and only allowing important information to appear in the output. The concept of matting can be employed to salient object detection where only the salient object information is allowed to pass to the output, while the unwanted background information is blocked. Matting can be performed by computing the matting components of an image, which can span any foreground and background components of an image. The fundamental problem in employing matting components to salient object detection is to identify and select only those matting components that belong to the foreground salient object and subsequently combine them to form the foreground object saliency.

The question of how to normalise and combine features (coming from different modalities) into a saliency map is vital for good performance. Harel et al. [50] obtained improved results by iteratively conditioning their final map in different cases. Prior works have also observed that different conditioning and combination strategies work for different types of images [65, 103]. For instance, the feature conditioning required for images having simple backgrounds will be different from those for images with cluttered backgrounds. *However, none of the previous methods have investigated image specific learning of combination schemes to increase generalisation.* A previous method [100] also acknowledged that performance of saliency methods varies with images and learned to combine the images so that



performance gaps between different models are minimised. However, it did not learn which normalisation and conditioning would suit a particular type of image and only relied on feature importance to minimize the gaps amongst model performance. The choice of normalisation and conditioning can be encoded as variables in a genetic algorithm framework along with feature importance, such that the parameters that maximize the agreement between model's output and the target output can be searched. In addition, the niching property of a Learning Classifier System (LCS) can be employed to perform image specific learning of feature combination schemes. An LCS can learn an image specific mapping of each image type to different feature combination strategies and autonomously identify the suited combination scheme to an image type.

Recent approaches in semi-supervised image classification and image retrieval require a small amount of labelled data and in contrast a large amount of unlabelled data [53, 156]. For visual saliency applications to aid computer vision approaches in such scenarios, there is a need to rank saliency results without the availability of ground truth data. This is also desirable in cases where a single *best performing saliency map* is preferred over all other maps for every image as the final output of the method. Moreover, in some saliency scenarios, it may be desirable to detect a noisy feature and exclude it from the combination to avoid unnecessary background noise. While there is not much previous work in this domain, a handful of related approaches have been introduced that attempt to select a single *best performing map* from a group of maps [27, 49, 56]. These approaches work well with a wide variety of images, however they can not handle difficult image types and cases where more than three features are involved. This is because these techniques select a map based on a few general properties associated with a good saliency map on the basis of prior knowledge. These simple observations do not generally hold on difficult image types (for visual examples, please see chapter 8 and chapter 9). Hence there is a need for devising cues that can determine the quality

of feature maps without the target label data. Such cues, once determined can be combined to form a robust quality measure of a feature/saliency map that can assist in dynamic selection of good quality features.

### 1.2.1 Why Genetic Algorithm?

Genetic Algorithm (GA) is an effective evolutionary technique for global optimisation and search. It is deemed suitable to address the global optimisation of important parameters of the visual attention model due to the following reasons:

1. Several efficient constrained optimisation techniques have been proposed in the literature to address problems with integer constraints [12, 16]. These techniques work well for a class of problems that are convex in nature, however when applied to non-convex problems, they might exclude the global optimum while constraining the search space. A class of such methods requires domain knowledge about the optimisation problem. While an alternative approach transforms the optimisation problem into a convex equivalent. Evolutionary computation techniques are competitive in large search spaces. Moreover, they do not require any prior knowledge of the search space and does not restrict the search space in the absence of constraints to be convex.
2. Evolutionary approaches such as genetic programming [81] and particle swarm optimisation [74] are also well suited for integer constraint based global optimisation and have proved valuable in addressing such problems. Several justifications for specifically considering a GA for this work are as follows:
  - (a) GAs provide a natural way to encode the real and integer parameters in a vector.

- (b) The integer handling can also be easily encoded in the GA and appropriate integer handling operators have been investigated in the past and are well-established [32].
- (c) GAs are easy to implement and have fewer parameters as compared with several EC techniques.

The main contribution of this work is the introduction of joint learning framework in the traditional computational model of visual attention, in order to increase its generalisability and not the comparison of the best suited optimisation technique.

### 1.2.2 Why Learning Classifier Systems?

Learning Classifier Systems (LCSs) are well known evolutionary machine learning algorithms that learn a population of classifier rules, which collectively address a problem. LCSs are a suitable method for learning feature combination in saliency detection for the following reasons:

1. LCSs are suited to image-specific learning of feature combination schemes, as they autonomously divide the search space into niches, so that a specific classifier in LCS can cover images of an identified type.
2. LCSs are highly flexible in terms of encoding of condition and action parts as compared with other rule based EC and non-EC techniques, which enables them to address multiple image types.
3. Due to their *if condition then action* rules, LCSs have the ability to autonomously learn and select the appropriate combination scheme based on the type of input image (*if image features then combination scheme*).

### 1.3 Thesis Statement

The proposed thesis is that artificial learning can increase the generalisability of the traditional model of visual attention by effective selection and optimal combination of features.

### 1.4 Thesis Goals

The overall goal of this thesis is to improve the generalisation of the traditional model of visual attention on unseen images and improve adaptability to data driven saliency applications. This goal is divided into the following sub-goals/objectives:

1. Develop a method to learn the various design choices of the computational model of visual attention for increased generalisation to unseen images. To achieve this goal the following research objectives have been set.
  - (a) Develop a new method that maximises the agreement between predicted saliency and the ground truth fixations for joint learning of feature importance and search for important parameters. The introduced method will be compared against the baseline method that employs fixed design parameters on the task of human eye fixation prediction.
2. Investigate novel features that seek to partition the image into foreground and background regions. These features will model saliency computation as a figure-ground segregation problem for accurate foreground background segmentation. To achieve this goal, the following research objective has been identified.
  - (a) Devise a new matting based feature that can mitigate the adverse affects of region based saliency computation such as inappropriate annotation and non-uniform saliency assignment.

Inappropriate object annotation and non-uniform saliency assignment are undesired characteristics of a saliency map. In the former, background regions are falsely highlighted, while in the latter, notably different saliency is assigned to various salient object regions. It is anticipated that the new feature will be able to uniformly highlight the salient object and generate saliency maps with minimal background noise. The proposed feature will be compared with well-known region based saliency methods on the task of salient object detection and segmentation.

3. Develop an approach that enables joint learning of feature related parameters *and* feature importance weights for the combination of complementary features for salient object detection. It is hypothesised that joint learning of feature related parameters by optimisation of an object detection based objective, the generalisability of the salient object detection method can be improved. To achieve this goal the following research objectives have been set.
  - (a) Develop a novel method for joint learning of important parameters by minimising the difference between predicted saliency and the ground truth segmentations for the task of salient object detection and segmentation.  
It is anticipated that the proposed method will obtain better features than features obtained through fixed parameters and combine them more efficiently than the baseline method.
4. Develop methods to autonomously identify a suitable combination scheme based on image type. It is anticipated that by learning multiple combination schemes and employing the best-suited scheme based on each autonomously identified image type will improve the generalisability of the traditional visual attention model on unseen image types. The following research objectives have been formulated to achieve these goals.

- (a) Develop a new multiple GA based approach for learning multiple feature combination schemes based on semi-autonomous identification of image types. The proposed method will be compared with classification based benchmark learning methods and other state-of-the-art methods on the task of salient object detection.
  - (b) Develop a novel computed action based Learning Classifier System (LCS) for learning multiple feature importance rules through autonomous identification of image types. It is anticipated that the LCS based technique will be able to learn maximally general feature importance rules by natural division of the images into niches. The developed method will be evaluated on the task of salient object detection using the standard evaluation procedures [3].
5. Investigate label free evaluation<sup>1</sup> of feature/saliency maps to select discriminative features based on their quality. It is hypothesised that quality based feature/saliency selection will improve the adaptability of the proposed approach to complementary feature selection and saliency aggregation tasks (for details of these tasks, please see the background chapter). The following research objective is defined to achieve this goal.
- (a) Develop unique cues to measure a feature/saliency map's quality without the knowledge of ground truth in order to autonomously identify discriminative features for appropriate feature combinations. It is envisioned that the developed system will be able to select features with high saliency prediction capability and deselect unimportant features during combination. The result will be an improved saliency output with improved adaptabil-

---

<sup>1</sup>Label free evaluation of feature/saliency maps refers to measuring the quality of maps without any ground truth data.

ity to tasks such as complementary feature selection and saliency aggregation. The introduced system will be evaluated by its performance comparison with well known saliency detection methods on the tasks of feature selection and saliency aggregation. Further evaluation of the proposed system will be conducted by comparing its salient object detection performance with the state-of-the-art salient object detection models.

Finally, these sub-goals/objectives will be brought together to analyse the effect of each contribution on the complete system.

## 1.5 Thesis Contributions

This thesis contributes to the following important issues in the field of machine vision in general and specifically in the field of salient object detection.

1. Joint optimisation of feature related parameters and feature importance weights was introduced to improve the saliency detection performance of the traditional visual attention model. Important parameters of the feature computation process and feature importance weights were learned by optimising a task specific objective function for human fixation prediction. By maximising the agreement between predicted saliency and the target (human fixations), the proposed GAOVSM method improved upon the performance of eight deterministic state-of-the-art saliency detection techniques on the task of human fixation prediction.
2. Spectral matting was employed for the first time in saliency prediction to combat the artifacts of region-based approaches for salient object detection. A novel saliency method for figure-ground segregation (termed as FGS) was introduced that employed matting components to construct smooth, uniform and accurate saliency maps. The

FGS method was able to overcome the artifacts of regions-based approaches by assigning uniform saliency to object regions while suppressing unwanted background information. As supported by the quantitative performance on salient object detection, the proposed FGS approach improved upon several state-of-the-art techniques.

3. Joint optimisation of feature computation parameters and feature importance weights was introduced for optimal combination of FGS with complementary features. Feature related parameters and their respective importance was learned at multiple segmentation thresholds by maximising the area under the precision-recall curve as an objective. The developed FGSopt method improved the object detection performance of the FGS technique and improved upon several state-of-the-art salient object detection models by considering the performance gaps amongst features.
4. *Semi-Autonomous identification of image type* was introduced to learn multiple feature combination schemes. Multiple combination schemes were learned from distinct image groups using multiple Genetic Algorithms (GAs). Images were placed into distinct groups using a semi-autonomous method that relied on their feature composition. By employing a suitable combination scheme for each unseen image type, the proposed image based GA (IGA) approach exhibited better generalisation as compared with a baseline GA that learned a single combination scheme. The IGA method also exhibited significantly better performance as compared with two classification based benchmark methods and three state-of-the-art models on the task of salient object detection.
5. Introduced *autonomous identification of image types* for learning multiple feature importance rules in order to increase the generalisability of the system on unseen image types. A supervised XCS based method was introduced that divided the search space into niches in



order to learn effective feature importance rules. This was achieved by employing a novel encoding scheme and a suitable action computation function. The proposed XCS based method improved upon the performance of the previously proposed multiple GA based method by obtaining a set of generalised feature importance rules.

6. Novel cues were established for dynamic feature selection in order to advance the current state of complementary feature selection and feature/saliency aggregation. Saliency quality measuring cues were introduced to seek discriminative features for appropriate combinations. Label free measurement of feature quality enabled the proposed feature selection method to improve upon the state-of-the-art in complementary feature selection and saliency aggregation. The proposed DFS based object detection method also improved upon seven state-of-the-art salient object detection methods.

## 1.6 Thesis Organisation

The rest of the thesis is organised as follows: chapter 2 presents the related work and background information. Chapter 3 to chapter 8 present major contributions of this thesis to achieve the corresponding goals listed out in this chapter. Figure 1.2 presents the general structure of these contributions chapters. Chapter 9 presents discussion on the interaction of the proposed approaches, while chapter 10 presents the main conclusions and highlight future directions.

Chapter 2, presents an introduction of the general structure of the traditional model of visual attention and gives an overview of the various available design choices. Important related work proposed for salient object detection is reviewed with important shortcomings highlighted, which form the motivation for the work presented in this thesis. Background information about the machine learning methods introduced in this thesis, i.e.

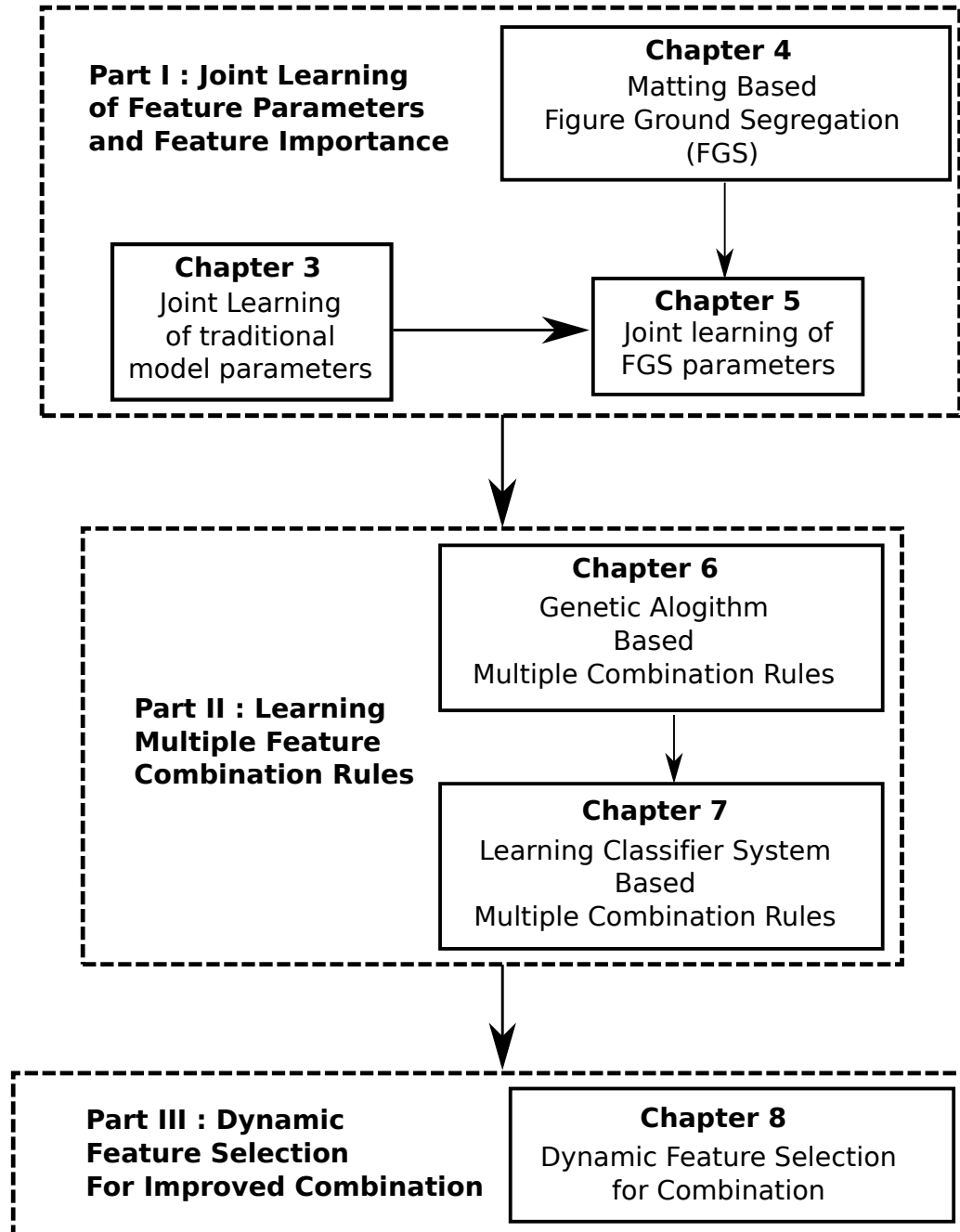


Figure 1.2: General structure of the contributions chapters.

Genetic Algorithms and Learning Classifier Systems are also presented.

In chapter 3, the traditional model of visual attention is equipped with the ability to learn important parameters through joint optimisation of feature related parameters and feature importance. By utilising the learned parameters, the resulting method is able to exhibit better generalisation as compared with the baseline counterpart and other benchmark methods.

The extended traditional model introduced before is not anticipated to work well on the task of salient object detection due to its simplistic features. The existing region-based feature computation approaches face inherent unwanted artifacts of region-based processing. A spectral matting based figure-ground segregation system (FGS) is introduced in chapter 4 to overcome the artifacts encountered by region-based salient object detection approaches. By suppressing the unwanted background information and uniform saliency assignment to object parts, the FGS approach improves upon several state-of-the-art salient object detection approaches.

In chapter 5, the generalisation performance of the FGS approach is improved by optimal combination of FGS with complementary features. The feature related parameters and importance are learned by considering the gaps between individual feature performance. The resulting method improves the generalisation performance of the FGS approach on unseen images for salient object detection. The improved figure-ground segregation approach also shows better performance as compared with several current state-of-the-art salient object techniques.

In chapter 6, multiple feature combination schemes each suited to a particular image type are introduced to further enhance generalisation of the saliency method. A semi-autonomous strategy is introduced to group images according to their types. By application of appropriate combination strategies to unseen image types, the resulting method improves the generalisability of a single combination scheme based method, while exhibiting better performance as compared with two classification based benchmark methods and several deterministic state-of-the-art models for

salient object detection.

In chapter 7, a fully autonomous grouping technique is introduced to overcome the limitations of the semi-autonomous approach of chapter 6 and to further enhance generalisation. The images are grouped into niches according to image type and multiple feature importance rules are learned each suited to a particular image type. By more accurate grouping of images and appropriate application of feature importance rules, the resulting method improves upon the performance of the semi-autonomous approach.

In chapter 8, a method is introduced to autonomously select the best feature maps for combination while neglecting undesired features on an image-by-image basis. To detect best performing maps and deselect the unwanted ones, novel feature quality measurement cues along with a clustering technique are introduced. The ability to judge features dynamically aids the benchmark methods to neglect unwanted redundant features and improve salient object detection performance.

Chapter 9 discusses the interconnections between the proposed approaches, the interaction of proposed systems in terms of a unified model and recommendations for future salient object detection models.

Chapter 10 presents the achieved objectives and major contributions of the thesis along with suggesting open questions for future research.

# Chapter 2

## Background

This chapter starts with an introduction to the general structure of the traditional model of visual attention and gives an overview of the available design choices. Section 2.2 introduces important methods proposed for the task of salient object detection, highlights their shortcomings, forming the motivation for the work presented in this thesis. Section 2.3 provides an overview of the machine learning methods adapted in this thesis to overcome the limitations of the previous methods by improving generalisation in salient object detection.

### 2.1 Computational Methods of Visual Attention

Inspired by the efficiency of the human visual attention system, researchers in computer vision, human-computer interaction, robotics and computer graphics are exhibiting increased interest in a mechanism that selects the most relevant information from a huge amount of input visual data. The objective of making such an approach is to improve vision systems and to understand human visual perception. The desired output for such mechanisms is a 2D map of the real world where a numerical value is associated with each location that is the likelihood of attending to it. Generally, these methods are similar in structure but have some variation in the details.

### 2.1.1 General Structure of Computational Models

The general structure followed by the majority of visual attention models is depicted in Figure 2.1. The model is adopted from Feature Integration Theory of Attention [130] and Guided Search model [143]. The first algorithmic implementation of the model appeared in the work of Koch and Ullman [79]. Since the first computational model of Koch and Ullman, many computational models have been introduced in the literature, many of which are extensions of the baseline Itti model [66]. In this chapter, the details of the specific model by Itti et al. [66] are discussed due to its validity as demonstrated by its applications for theoretical understanding of human attention and its practical applications for attention based systems. Furthermore, it has been adapted by many previous studies for understanding human selective attention [65, 107, 134, 153, 154].

The basic idea of the model is to find the regions of the input that differ from their surroundings. The computational model of attention employs three basic features; colour, intensity and orientation. By choosing these three particular features, the contrast is restricted to these domains.

The original image  $I$  is subjected to linear filtering and sub-sampling to obtain an image pyramid with  $l$  levels. An image pyramid consists of a number of filtered images where the number is defined by the levels  $l$ . Each level is obtained by convolving the input image received from the previous level by convolution with a separable Gaussian filter and decimation by a factor of two. Each map of the intensity pyramid is subjected to the following operation to obtain an intensity pyramid  $M_I(l)$

$$M_I = \text{mean}(R, G, B), \quad (2.1)$$

where  $R$ ,  $G$ , and  $B$  are red, green and blue values of the colour image.

The idea of mapping the colour to opponency axes is to cover the entire visible light [58], while the specific transformation is to eliminate the influence of illumination. For this purpose, red-green (RG)  $M_{RG}(l)$  and blue yellow (BY)  $M_{BY}(l)$  pyramids are obtained by subjecting each level of the

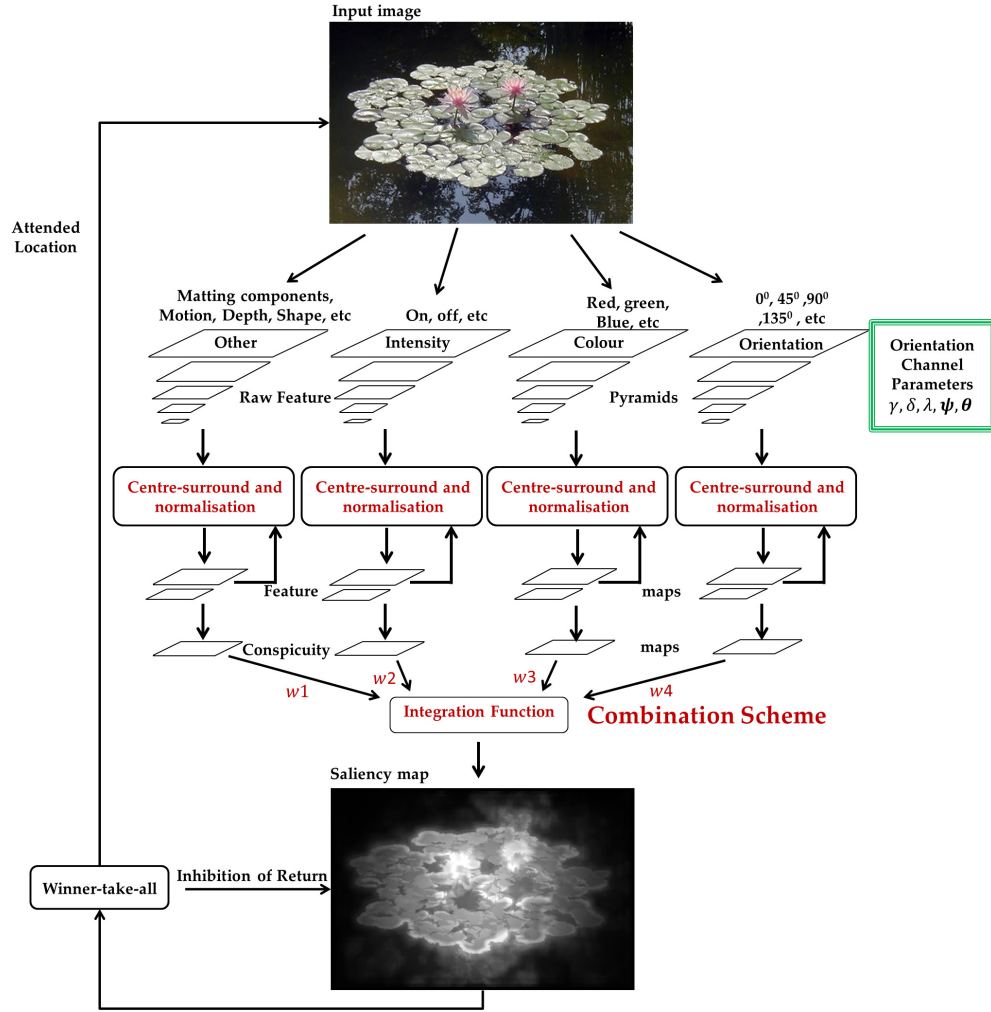


Figure 2.1: The general feature-based computational model of visual attention.

image pyramid to the following colour opponency transformation.

$$M_{RG} = \frac{R - G}{\max(R, G, B)}, M_{BY} = \frac{B - \min(R, G)}{\max(R, G, B)} \quad (2.2)$$

Local orientation pyramids  $M_\theta(l)$  are obtained by Gabor filtering of the intensity pyramid levels using the convolution operation:

$$M_\theta(l) = \|M_I(l) * G_{\psi_1}(\theta)\| + \|M_I(l) * G_{\psi_2}(\theta)\|, \quad (2.3)$$

where the Gabor filter is expressed as:

$$\begin{aligned} G_\psi(x, y, \theta) &= \exp\left(-\frac{\hat{x}^2 + \gamma^2 \hat{y}^2}{2\delta^2}\right) \cos\left(2\pi \frac{\hat{x}}{\lambda} + \psi\right) \\ \hat{x} &= x \cos \theta + y \sin \theta \\ \hat{y} &= -x \sin \theta + y \cos \theta. \end{aligned} \quad (2.4)$$

Here in (2.4),  $\gamma$  represents the aspect ratio,  $\delta$  represents the standard deviation, wavelength of the filter is  $\lambda$ , orientation  $\theta$  represents four local orientations  $\theta_1, \theta_2, \theta_3, \theta_4$ , and phase  $\psi$  represents the two phases  $\psi_1, \psi_2$ .

Figure 2.2 illustrates the effects of varying the wavelength ( $\lambda$ ) and elongation ( $\gamma$ ) on the Gabor filter kernel, while Figure 2.3 depicts the effect of varying orientations ( $\theta$ ) and phase ( $\psi$ ). All other parameters are kept fixed. It can be observed that increasing  $\gamma$  increases the ellipticity and support of the Gabor function, while increasing the wavelength has an obvious effect on the visible parallel excitatory and inhibitory stripe zones. The parameter  $\theta$  changes the orientation of the normal to the parallel stripes of the Gabor function, which in turn affects the specific orientations captured during the filtering operation. The change in  $\psi$  correspond to symmetric and anti-symmetric functions imitating the “centre-on” and “centre-off” functions of the human visual system as depicted in Figure 2.3.

Centre-surround visual receptive fields are implemented by across scale subtraction ( $\ominus$ ) between centre (c) and surround (s) levels to obtain feature maps. Levels in the pyramids are designated as centre (c) and surround



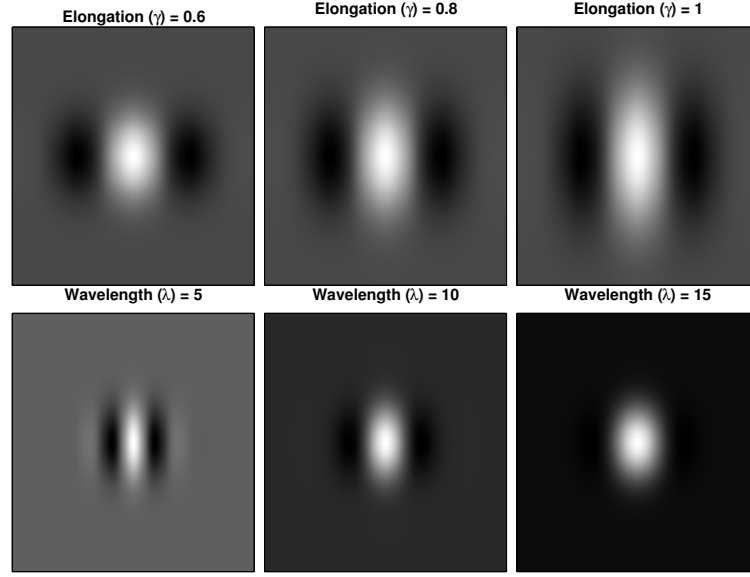


Figure 2.2: Effects of the wavelength ( $\lambda$ ) and elongation ( $\gamma$ ) on the support of the Gabor filter kernel.

levels ( $s$ ) based on their position in the pyramid. The centre-surround operations capture the within contrast of the maps by comparing the average value of the centre region ( $c$ ) to the average value of the surround region ( $s$ )

$$F_{C,I,O} = \mathcal{N}(|M(c) \ominus M(s)|),$$

$$\forall M(l) \in \{M_I(l)\} \cup \{M_C(l)\} \cup \{M_O(l)\}, \quad (2.5)$$

where  $M_C(l) = \{M_{RG}(l)\} \cup \{M_{BY}(l)\}$ ,  $M_O(l) = \{M_{\theta_{1v}}(l)\} \cup \{M_{\theta_{2v}}(l)\} \cup \{M_{\theta_{3v}}(l)\} \cup \{M_{\theta_{4v}}(l)\}$  and  $\mathcal{N}(\cdot)$  is a simple normalisation operator used to scale the values in feature maps to a fixed range  $[0, M]$ . Other complex forms of normalisation operations that are commonly employed to eliminate any differences between non comparable modalities (extracted using different extraction mechanisms) are discussed below.

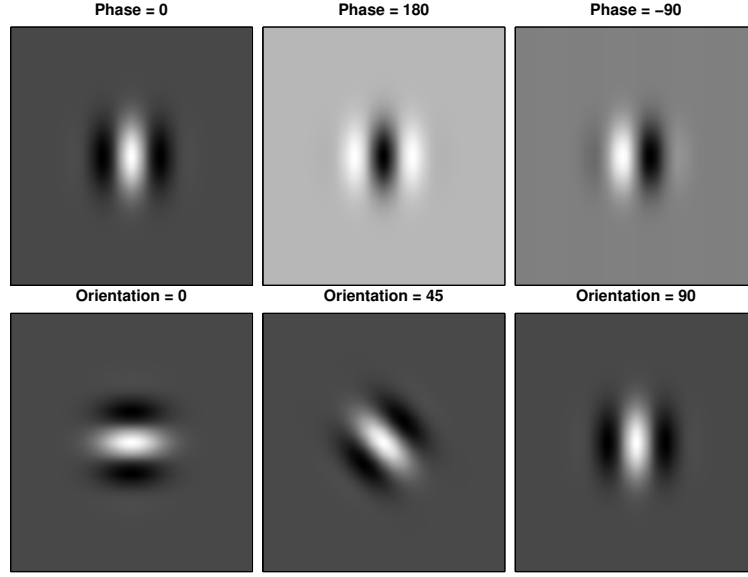


Figure 2.3: Effect of varying orientations and phase on the Gabor filter kernel.

For each feature from (2.5), the feature maps are combined using across-scale addition ( $\oplus$ ), across centre-surround (c-s) scales and the results are normalized again to the same range:

$$F'_{C,I,O} = \mathcal{N}(F_{C,I,O}(c) \oplus F_{C,I,O}(s)). \quad (2.6)$$

Conspicuity maps are then computed by summing all the sub-features for each general feature and subsequent normalization.

$$C_I = F_I \quad (2.7)$$

$$C_C = \mathcal{N}(F_{RG} + F_{BY}) \quad (2.8)$$

$$C_O = \mathcal{N}(F_{\theta 1} + F_{\theta 2} + F_{\theta 3} + F_{\theta 4}) \quad (2.9)$$

Finally, all conspicuity maps are weighted and combined linearly to produce the saliency map  $S$ .

$$S = \frac{1}{3}(w_I C_I + w_C C_C + w_O C_O) \quad (2.10)$$

The saliency map  $S$  is a gray-scale image that depicts the level of importance of a pixel by its brightness.

The saliency map is the final output of most computational models, however, some applications require attending to more than one salient region to mimic human saccades. The salient regions are the local maxima in the saliency map and are attended to in a sequential fashion using a *winner-take-all* (WTA) approach [78]. A notion of *inhibition of return* (IOR) [115] is implemented to ensure that the focus of attention does not remain fixated at the most salient location. As this thesis is concerned with salient object detection in static images, the final saliency output of the saliency detection system is of concern but any subsequent processes of attending to a location and generation of a saccades are out of the scope.

An important aspect in any attention system is determining the relative importance of the various features during fusion. Usually a weighting function is applied to each map before fusion. Additionally, as the maps come from non-comparable modalities, it is important to eliminate any differences before their combination. Finally, the choice of integration of the maps to form the final saliency map carries utmost importance. It is worth highlighting that the issues concerning the relative importance of features and choice of integration are not considered by the model of Itti et al. [66]. However they greatly influence the performance of the visual attention system. The following sections provide details regarding these processes.

### Feature Weighting

The weight assigned to a feature quantifies its relative importance in predicting saliency. Each feature map is multiplied by its corresponding weight during the combination process to control its relative contribution to the final saliency map. These weights represent the top-down information based on prior information about the features that are important for a given task. A class of methods try to find suitable values for this top-down

information in order to optimise the behaviour of the saliency method.

### Normalisation

As various feature maps come from different modalities and dynamic ranges, they are not immediately commensurate. Therefore, a scheme is required that promotes feature maps with a small number of strong peaks of activity (“odd man out”), while suppresses maps exhibiting comparable peak responses at numerous locations in the visual scene. This is achieved through normalisation *normalisation*<sup>1</sup>, which is a crucial step in feature combination. Some normalisation functions reported in the literature are global and iterative [65]; identity, exponential and logarithmic [103], and a selection can be expressed as follows:

$$\left\{ x, \exp(x), \frac{-1}{\log(x)}, \frac{1}{1 + \exp(-x)}, x(M - \bar{m}), x + x * \mathcal{DoG} \right\}, \quad (2.11)$$

where  $\mathcal{DoG}$  is the difference of Gaussian filter and  $*$  denotes convolution.  $M$  is the global maximum of the feature map and  $\bar{m}$  represents the local minimum of the feature map.

Figure 2.4 demonstrates the importance of applying appropriate normalisation scheme as per the input type. Two different examples are selected for illustration purposes. The first row presents a synthetic feature map containing numerous strong peaks, while the second row features a real retina image considered here as a feature map with intensity representing saliency in both cases. The second column in Figure 2.4 plots the 3D profiles of the input feature maps, while the third and fourth columns present the profiles after subsequent normalisations. The applied normalisation scheme in Figure 2.4 is  $\mathcal{DoG}$ . The difference of Gaussians  $\mathcal{DoG}$  is an approximation of Laplacian of Gaussians and allows for relatively simpler and faster computation [82].

---

<sup>1</sup>The normalisation function used in the context of saliency detection methods is more general than the linear scaling understood for the conventional normalisation function.

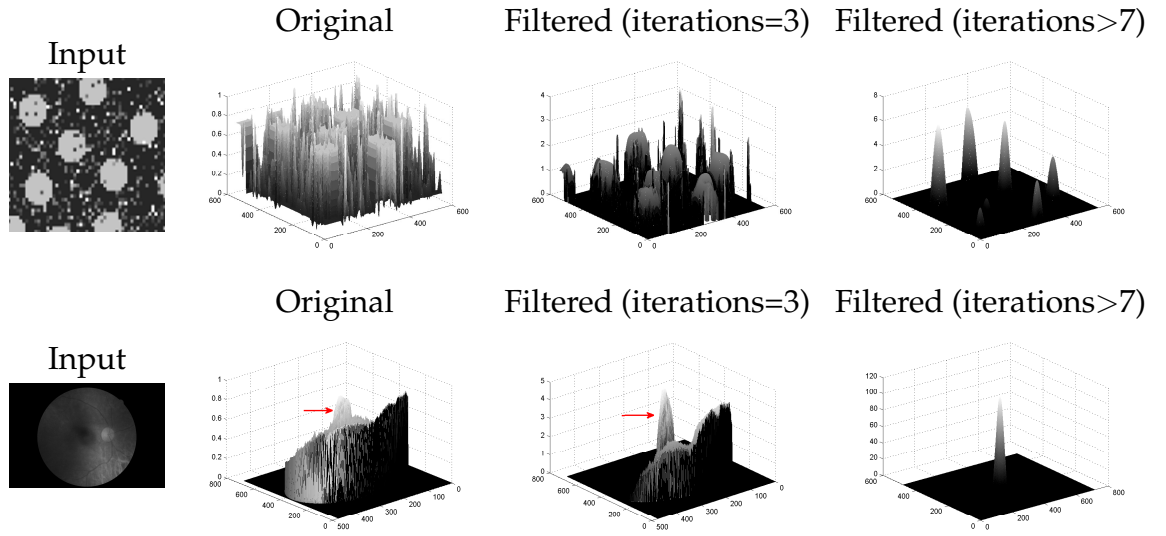


Figure 2.4: Importance of the normalisation scheme suited to input type. The red arrows are drawn to highlight the strong activation peaks. Plots in column 2 and 3 of the second row are rotated for better visualisation of the strong peaks.

It can be observed that for the first row map, the filtering operation uniformly suppresses all the activation peaks up to three iterations. Going beyond three normalisation iterations, a few peaks are disproportionately suppressed, while several others appear as strong peaks in the output. This behaviour of the applied normalisation scheme suggests that it is not suited for maps that have several equal strong peaks in their landscape. In contrast, the 3D profile of the retina map (in the second row) appears to have one strong activation peak surrounded by several weaker ones that constitute the noise in the map. Conversely, on the retina image, the  $\mathcal{DoG}$  based normalisation is highly effective in suppressing the noisy peaks and aid the strong peak to stand out. Notably, increasing the number of iterative normalisations for the second row map results in complete suppression of the unwanted noise, while strengthening of the strong peak.

The illustration of Figure 2.4, stresses the importance application of

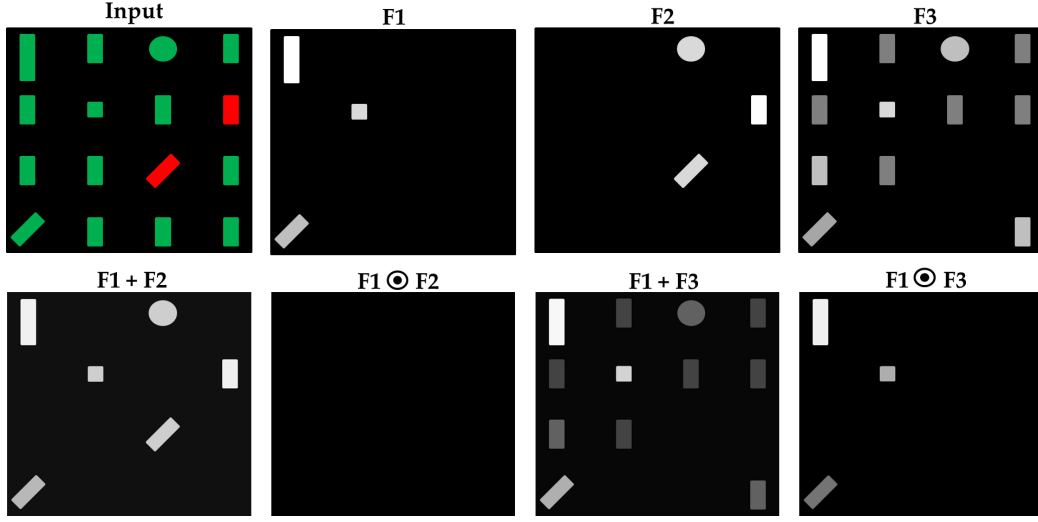


Figure 2.5: Choice of integration operation depending upon the types of input features.

appropriate normalisation scheme depending upon the input map type. No prior investigation has been conducted into which of the schemes is best suited to various types of images and settings.

### Integration

The mathematical operation that combines the feature maps to produce the final saliency output is termed the integration function. Addition of feature maps has been typical for most methods [13, 72, 153, 154]. However, Klein et al. used element-wise multiplication [77] and Gazit et al. used the harmonic mean [52] to integrate the feature maps. These integration functions can be expressed as follows:

$$\left\{ \sum_{i=1}^n x_i, \prod_{i=1}^n x_i, \left( \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \right)^{-1} \right\}. \quad (2.12)$$

The simple summation of the features is not necessarily the best choice

as a high value falsely assigned to a background pixel by a feature can corrupt the final saliency response. Element-wise multiplication may seem a better choice as multiple feature maps have to predict a pixel as salient for it to be considered salient in the final response. However, this approach has the potential of producing saliency maps with low recall. The harmonic mean of maps is only suited to features containing outliers; meaning background pixels falsely predicted as salient. Conversely, on features with fewer outliers, it may negatively affect the strong salient peaks in the feature maps.

As an example, Figure 2.5 illustrates the importance of appropriate integration scheme selection as per map type. The input stimulus is a synthetic psychological pattern employed in pop-out search tasks. The objects having different shapes orientation or colour are likely to pop-out in a search task, while the similar shape and colour objects are distractors. Feature maps F1, F2 and F3 are created to illustrate the importance of choice of integration based on feature type. Several feature combinations are depicted in the second row of Figure 2.5.

It can be observed that the addition of F1 and F2 results in a good output map. The reason for this result is that F1 and F2 are highly uncorrelated capturing totally different subsets of the input. Conversely, the element-wise multiplication of F1 and F2 results in an all zeros, as they do not have anything in common. In contrast the addition of F1 and F3 results in including several distractors, while their product results in a better map containing a subset of the target objects. The reason for this result is that F1 and F3 are correlated in terms of including the same target objects.

### 2.1.2 Choice of Parameters to be Optimised

The performance of computational models of visual attention heavily depends upon tuning several design parameters discussed in the previous sections. A few examples are:

1. Choosing the type (Gaussian or Laplacian) and levels of image pyramids.
2. Optimal modelling of visual receptive fields, difference of Gaussian filters (for non oriented features) or Gabor filters (for oriented features) [65].
3. Appropriate application of normalisation and integration schemes, depending upon the input type.
4. Optimal weighting parameters.

Each of these design choices have a profound effect on the final saliency output of the computational model. They are also interconnected and are affected by each other [75]. Moreover, it can be intuitively realised that the quality of computed features will affect the relative feature importance weights.

Despite the introduction of methods to learn feature weightings (section 2.2.1), no previous method has been discovered that can jointly optimise the important parameters and the feature importance weights. Additionally, no prior effort has been made to identify appropriate normalisation and integration schemes, depending upon the type of input imagery.

After discussing the general structure of computational model of visual attention and the need for effective search of design choices, the next section will discuss the task of salient object detection. Important computational methods built for the task of salient object detection as reported in the literature will be discussed.

## 2.2 Salient Object Detection

Most computational methods of attention were built for human fixation prediction inspired by human perception [13, 72, 106]. Recently, visual saliency has attracted much computer vision research, giving rise to a new





Figure 2.6: Difference between salient object detection and figure-ground segmentation. For the first two images, a salient detection algorithm is expected to segment out the red pin in the first image and only the pink region in the second image (both depicted by blue contours). Contrarily, a figure-ground segmentation is likely to segment the regions inside the green contours as foreground. On the physiological pattern in the right corner, the figures should be perceived as all the letters, while a salient object detection algorithm will only consider L as important.

subdomain known as salient object detection. In salient object detection, the task is to detect the salient, attention grabbing object(s) in a scene. In essence, salient object detection is similar to the problem of figure-ground segmentation as both seek to separate foreground objects from background. The subtle differences between the two fields are depicted in Figure 2.6. Salient object detection also differs from the traditional segmentation problem, where the task is to partition the image into perceptually homogeneous regions. It is a difficult problem in computer vision as natural scenes can have objects with cluttered backgrounds (making it difficult to distinguish the object from background) and can contain multiple ambiguous salient objects. The next section explores a few important approaches to salient object detection, groups them into different categories and highlights their shortcomings paving the way for the methods introduced in this work.

### 2.2.1 Important Computational Methods for Salient Object Detection

The methods introduced in the literature for the task of salient object detection can be generally categorised as follows:

#### Deterministic Methods

Based on their computation style, the methods belonging to this domain can be further categorised as block-based and region-based methods. Block-based methods employ image blocks as their visual subsets. These are usually the early methods in salient object detection history, while region-based approaches employ the increasingly popular superpixel algorithms to obtain their visual sets.

A few block based methods have been introduced in the past that utilize cues extracted from blocks of images. Goferman et al. [46] introduced a method termed as CA<sup>2</sup> that used the distinctiveness of image patches as compared to their most similar patches, while considering the spatial distances between patches. Achanta and colleagues introduced three influential block based methods for salient object detection. The first method termed as AC [2] employed local contrast of image blocks in terms of colour and luminance features to determine its saliency. The second method, termed as FTS [3], introduced a frequency-tuned approach that measures saliency as the contrast between the filtered features and the arithmetic mean of features by treating a pixel as the centre and the whole image as the surround. The third method by Achanta and colleagues [5], termed as MSSS, replaces the centre-surround approach of FTS that treats the whole image as the surround by a symmetric surround with respect to each pixel. In comparison to FTS, the symmetric surround approach of MSSS allows more local surround at the image borders. Margolin et

---

<sup>2</sup>The techniques that follow will be utilised as benchmark approaches so will be termed with short names for later reuse in the results chapters.

al. [101] introduced a method, termed as PCA, to measure the uniqueness of a block by its distance from the average block in the high-dimensional space. The unique or more salient blocks must be more scattered than the non-distinct patches.

For block based approaches, the inter-block contrast is very low when operating inside the salient object contours and comparatively high at salient object edges. Therefore a common shortcoming of all of these block based methods is that high contrast edges are usually highlighted instead of the salient object. A plethora of region based approaches have recently been proposed to overcome these limitations of the block based methods.

The regional contrast method of Ming et al. [28] (termed as RC), measures the saliency of a region as its global contrast to all other regions in the image. Hence, regions having high global contrast with respect to other regions will be assigned high saliency values according to RC [28]. In contrast to the global contrast measure of Yan et al. [147] (termed as HS) employed local region-based contrast to capture a region's saliency, where each region is an outcome of a watershed-like operation on the input image [147]. The approach of Zhu et al. [157] termed as MND captured the perceptual similarity of regions by combining the benefits of both local and global saliency. MND clusters similar regions to identify the distinct regions that are likely to be salient object parts using its global saliency, while it uses a local saliency cue to highlight regions that stand out from their surroundings.

The work of Yang et al. [148] termed as MR, proposed a two stage saliency computation framework by manifold ranking of regions using an undirected weighted graph. During the first stage, a saliency score for each region is computed based on its relevance from the pseudo-background regions. Based on the saliency scores for regions from the first stage of processing, foreground regions are estimated and used as queries to further refine the saliency scores of regions. Similar to MR, the work of Li et al. [91] termed as DSR, formulated the saliency computation prob-

lem for a region as the dense and sparse reconstruction error from the pseudo-background regions. The reconstruction errors are propagated on to each pixel at multiple scales. Finally, multi-scale information is fused together to form the final saliency map using Bayesian inference. On similar grounds the approach of Jiang et al [67] termed as MC computes the saliency of a region as the absorbed time from the transient node (i.e. regions belonging to the image centre) to the absorbing nodes (i.e. regions along the image border).

Despite the recent popularity and success of the region based methods, the shortcomings of these methods are that they fail to uniformly highlight the salient object and can completely miss parts of the salient objects in the final saliency map. These shortcomings are generally a consequence of the hard decision boundaries of superpixels<sup>3</sup>, which do not necessarily respect object boundaries. The recent approach of Ming et al. [26] termed as SIA, attempted to overcome the problem of non-uniform saliency assignment to salient objects by replacing hard boundaries by soft image abstractions. However, the response of SIA is still subject to non-uniform saliency assignment when the regions with soft image abstractions fail to respect object boundaries.

The shortcomings of the region-based approaches are depicted in Figure 2.7. The rationale for choosing the first test image is that the foreground object shares colour and intensity features with numerous background regions, which makes the discrimination between foreground and background regions difficult when the regions are predetermined as a pre-processing step. The first row presents the various segmentation levels employed by the multi-level saliency computation approach of DRFI [68]. It can be observed that the object starts to appear coherently with the de-

---

<sup>3</sup>The term superpixel is widely accepted in the segmentation literature to define a collection of similar pixels in the image. Lately, it has been rigorously employed by salient object detection methods as a pre-processing step to segment the image into perceptually homogeneous regions.

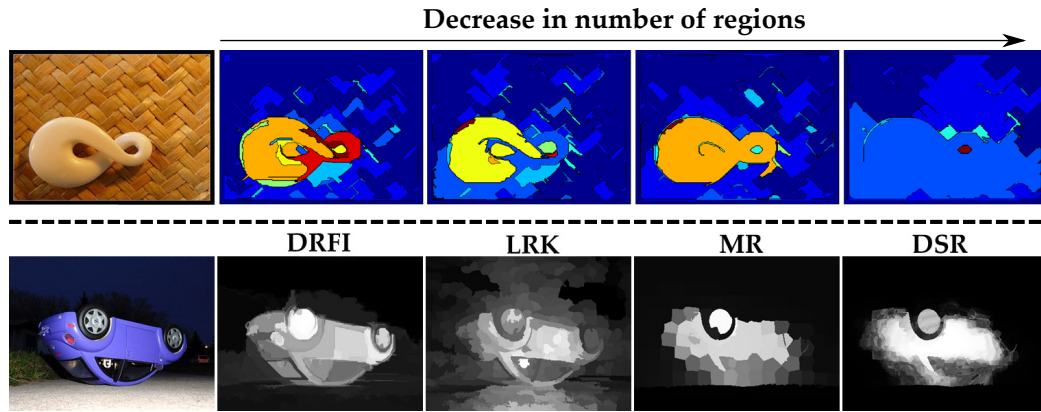


Figure 2.7: The artifacts of region based processing. The first row shows the effect of various segmentation levels employed by the multi-level saliency response of DRFI. The second row presents outputs of four region-based benchmark methods, including DRFI [68], LRK [123], MR [148] and DSR [91]. The example images presented here are selected to best depict the artifacts of a region based pre-processing step commonly employed by benchmark methods. Please refer to text for details.

crease in the number of regions, however, further decrease in the number of regions completely loses the object. Such maps, as depicted in the first row, column four of Figure 2.7, can corrupt the overall output response of DRFI. The second image is chosen as it contains an object composed of various regions with different colour and textural properties (e.g. windows, doors and tyres of the beetle car). The second row shows a representative car image along with the characteristic region-based response of four benchmark methods. It can be observed that LRK and MR literally break the object (i.e. the car) into pieces, resulting in non-uniform saliency assignment. The saliency outputs of MR and DSR also fail to capture part of the car and miss the front tyre.

A major issue with deterministic methods is that they employ ad hoc strategies at various stages of feature combination. Representative examples include the complementary local and global saliency maps of the MND method, where element-wise multiplication of maps is employed to obtain the final saliency map. The element-wise multiplication scheme ensures that no background information is falsely included in the final response at the cost of saliency maps with non-uniformly highlighted salient objects and at times partly highlighted salient objects. Another instance of the use of element-wise multiplication of maps is reported at the first stage of MR, where the saliency maps obtained from each of the four pseudo-background regions are combined through element-wise product. This process also leads to loss of salient information, when some part of the salient object overlaps with one of the four background regions.

### **Learning Based Methods**

There is an abundance of deterministic methods for salient object detection, which employ ad hoc feature weighting, normalisation and integration strategies. However, there are also limited number of studies that explore the potential of supervised learning to determine better feature combination strategies. Several methods that employ supervised learning

for feature combination for salient object detection and are directly related to the proposed work of this thesis are discussed here. These methods are used for benchmarking and comparison purposes in the result chapters of this thesis.

The method of Judd et al. [72] termed as LSVM used a support vector machine (SVM) to train their method for learning feature importance weights. Linear kernel was adopted in their method due to its superior performance as compared with non-linear counterparts. The drawback of the SVM based methods is that they can only learn feature weights as the joint learning of feature related parameters with feature weights can not be easily incorporated in their objective functions.

The recent work of Jiang [68] termed as DRFI performs multiple segmentations of the input image and learns a regressor that directly maps the raw features for each region to a saliency score. The learned regression method is then employed at the test stage to predict the saliency scores for each region and the final saliency is computed by fusing saliency maps computed over multiple segmentations. DRFI exhibits robust performance on multiple datasets and all salient object detection benchmarks at the cost of computing an immense 93-dimensional feature vector for each image adding to its computational time and limiting scalability. Although DRFI exhibits robust saliency response in general, it can contain artifacts of non-uniform saliency assignment, which is attributed to the regional contrast feature of DRFI (see Figure 2.7). In scenarios where feature performance is heavily dependent upon important feature related parameters, joint optimisation of such parameters can not be easily incorporated into the automatic feature integration process of DRFI. Additionally, on difficult scenes with cluttered background and multiple salient objects, the multi-level saliency computation stage of DRFI can include redundant saliency information during fusion, resulting in an undesired response.

The work of Shen and Wu [123], termed as LRK, combines low and high level features into a feature matrix that is decomposed into two parts,

the low-rank matrix and the sparse matrix with the assumption that non-salient background information will be captured by the low-rank matrix, whereas the salient regions are indicated by the sparse noise. Although, LRK attempts to separate salient regions from the background information, the low rank information can not be effectively separated from the sparse information in general and background information is included in the characteristic response of LRK.

In summary, the learning approaches do not investigate the application of appropriate normalisation and integration schemes as per input type. Moreover, the entirety of learning based approaches assume that a single learned strategy can suffice for all different types of image classes.

### **Feature Selection Based Methods**

Prior works have proposed complementary saliency features for the task of salient object detection [27, 49, 56, 69]. A subset of such methods that compute complementary saliency maps choose an individual (best performing) map to act as the final saliency output on an image-by-image basis. The process of finding and selecting the best feature map for a particular scene is termed as complementary feature selection. The few studies that have explored complementary feature selection for improving salient object detection are briefly reviewed here.

Gopalakrishnan et al. [49] modelled the distributions of colour and orientation to construct two complementary saliency features, namely, colour saliency framework (CSF) and orientation saliency framework (OSF). They proposed a new measure termed as the saliency index (SI), which selects the feature with the lowest variance in the spatial domain as the final output of their saliency method. SI assigns a usefulness measure to a feature map  $F$  depending upon the compactness and the connectedness of its segmented output. The compactness cue provides a measure of spatial



variance  $SV_{\mathbf{F}}$  for  $\mathbf{F}$  defined by

$$\begin{aligned} SV_{\mathbf{F}} &= SV_{\mathbf{F}}^{p_1} + SV_{\mathbf{F}}^{p_2} \\ &= \frac{\sum_{\mathbf{p}} (p_1 - \mu_{p_1})^2 \cdot \mathbf{F}(\mathbf{p})}{\sum_{\mathbf{p}} \mathbf{F}(\mathbf{p})} + \frac{\sum_{\mathbf{p}} (p_2 - \mu_{p_2})^2 \cdot \mathbf{F}(\mathbf{p})}{\sum_{\mathbf{p}} \mathbf{F}(\mathbf{p})}, \end{aligned} \quad (2.13)$$

where  $\mathbf{p}$  is a vector of dimensions  $l$  ( $l = 2$  here) used for indexing a feature map<sup>4</sup>, and  $SV_{\mathbf{F}}^{p_1}, SV_{\mathbf{F}}^{p_2}$  are the spatial variances and  $\mu_{p_1}, \mu_{p_2}$  are the spatial means for  $\mathbf{F}$  in the  $p_1$  and  $p_2$  directions, respectively.

The connectedness cue  $C_{\hat{\mathbf{F}}}$  for a segmented feature map  $\hat{\mathbf{F}}$  measures the average number of non-zero pixels in a neighbourhood

$$C_{\hat{\mathbf{F}}} = \frac{\sum_{\mathbf{p} \in \mathcal{P}} \sum_{\mathbf{p}' \in \mathcal{N}_{\mathbf{p}}} \mathbf{1}_{\mathcal{P}}(\mathbf{p}')}{|\mathcal{P}|}, \quad (2.14)$$

where  $\mathbf{1}_{\mathcal{P}}(\mathbf{p}) = \begin{cases} 1 & \text{if } \mathbf{p} \in \mathcal{P} \\ 0 & \text{if } \mathbf{p} \notin \mathcal{P}, \end{cases}$

where  $\mathcal{P}$  is a set of coordinates of non-zero pixels in  $\hat{\mathbf{F}}$ ,  $|\mathcal{P}|$  is the cardinality of the set  $\mathcal{P}$  and  $\mathcal{N}_{\mathbf{p}}$  is the neighbourhood of  $\mathbf{p}$ .

The saliency index ( $SI$ ) for feature map, according to Gopalakrishnan et al. [49] is given as:

$$SI = \frac{C_{\hat{\mathbf{F}}}}{SV_{\mathbf{F}}}. \quad (2.15)$$

The method is termed as SISal for the rest of the thesis. SISal performs good complementary feature selection using the SI measure and shows competitive performance on a benchmark dataset [94]. However its selection measure is biased towards compact orientation maps as compared with spatially distributed colour saliency maps [49].

Recently Cheng et al. proposed the complementary global saliency cues of colour spatial distribution (CSD) and global uniqueness (GU), which

---

<sup>4</sup>For consistency, as an abuse of notation, we use  $\mathbf{p}$  for indexing both vectors and images.

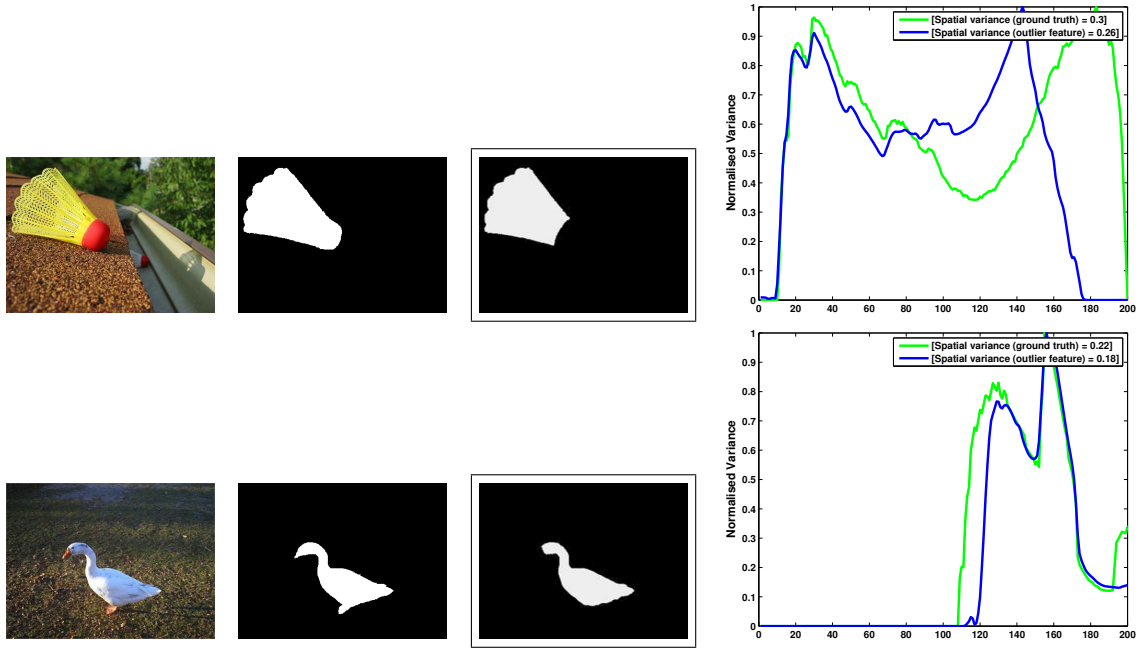


Figure 2.8: Shortcomings of the previous feature selection approaches, namely SISal and GC. Both the SISal and GC methods prefer features that are compact and are less varied in the spatial domain. From left to right: input image, ground truth (GT), an outlier feature and normalised spatial variance of the features. The GT is considered here as an input feature to test the feature selection performance of the methods. The boxes depict the selected features by both SISal and GC. The spatial variances of the features reveal that both the SISal and GC methods falsely select the outlier features (instead of the GT), due to the more compact nature of the outlier features.

were computed using soft image abstractions instead of hard segmentations [27]. Following the work of Gopalakrishnan et al. [49], the global cues (GC) method of Cheng et al. employed the compactness measure (termed as SV) of [49] to select one of its complementary features on an individual image basis. They achieved promising results on the FT benchmark [3]. However, a few cases are falsely predicted as the SI measure is focused on the compactness of the features irrespective of their salient object prediction capability.

Both SISal and GC heavily rely the compactness of features as a selection criteria. The shortcoming of the compactness measure is illustrated by a simple example in Figure 2.8. Two features for each of the representative example are shown. The second column includes the ground truth for the input images, which are considered here as features to test the feature selection capability of the compactness measure. The maps in the third column are representative outlier features that capture part of the salient objects. The last column shows the profiles of spatial variance (computed according to (2.14)) for both features. When SISal and GC were tested on these maps, they both selected the outlier features as their output, due to their high compactness. This result is explained by the spatial variance profiles of the features, which show that the ground truth feature is more spatially varied for both the cases, implying less compactness.

Hu et al. [56], introduced a measure, termed as CSI, based on spatial compactness (SC) and saliency density ( $D_{\text{saliency}}$ ) to measure the usefulness of a feature for dynamically selecting features on an image-by-image basis. The composite saliency indicator (CSI) [56] measure employed the spatial compactness (SC) of the salient region and the density of the salient region of a feature map to judge its usefulness. Salient points are first extracted from the salient region as described by a feature map. Afterwards the convex hull of these salient points is constructed. The area enclosed by the convex hull of the salient regions gives SC.

Saliency density  $D_{\text{saliency}}$  for a feature map  $F$  is computed as:

$$D_{\text{saliency}} = \frac{\sum_{p \in S} \frac{\sum_{q \in \mathcal{N}_p} |F(p) - F(q)|}{|\mathcal{N}_p|}}{|S|}, \quad (2.16)$$

where  $F(p)$  is the intensity at location  $p$ ,  $\mathcal{N}_p$  is the set of neighbouring points of  $p$  and  $S$  is the set of salient points. In (2.16),  $q$  is an index vector similar to  $p$  used for the neighbouring locations of  $p$ . The features are then selected based on the cues  $SC$  and  $D_{\text{saliency}}$ .

The evaluation process employed by Hu et al. [56] is unique as compared with the standard salient object detection evaluation procedure. To evaluate the final saliency output of the method, the polygon obtained by the salient points of the final saliency output is employed and only a visual comparison with the state-of-the-art is presented. Results from a user study are also reported to compare the performance of methods. The process of CSI based feature selection shows good performance in general, however, the process is highly sensitive to the polygon obtained by the convex hull and false inclusion of non-salient points can negatively affect the salient feature selection performance of the CSI measure.

CSI is more sophisticated than SI as it assigns a qualitative measure to the features rather than judging them with respect to their compactness only. However, the selection of maps based on the area of the convex hull makes CSI unrealistic on scenes containing large salient objects.

### Saliency Aggregation Based Methods

Saliency aggregation is the task of combining the best response of saliency methods. The need for saliency aggregation arises due to the fact that different saliency methods are tuned to capture different aspects of saliency for different images. From this arises the need to combine the better aspects of various saliency methods for a more general saliency system. Recently, a data-driven approach for aggregating the saliency outputs of state-of-the-art saliency methods was presented in [100]. Three different

methods were proposed for combining individual saliency maps by minimizing the performance gaps. The details of only one of their methods called the pixel-wise saliency aggregation (termed as PW), which employs logistic regression is described here as it is the only method used in subsequent investigation. They aggregate the results of various saliency methods pixel-wise by learning the appropriate weights (combination) using training data. The PW method is reported to outperform all the individual saliency methods. However, low performing methods degrade the overall results more than they aid in improving.

The approach of [68] also involve saliency aggregation to combine its multi-level saliency maps in order to obtain the final saliency map. DRFI employed a least square estimator to learn the weights for the multi-level saliency maps to perform saliency aggregation. Although, the DRFI method produces state-of-the-art results on benchmark datasets, the multi-level saliency fusion process is shown to degrade the overall performance of the system often by including unnecessary (redundant) saliency maps urging the need for selecting only the appropriate maps for combination (see Figure 2.7).

## 2.3 Machine Learning Methods

Numerous evolutionary approaches have been employed for computer vision tasks including genetic programming [7, 80] and particle swarm optimisation [95]. A plethora of EC techniques have also been applied to feature selection for classification tasks using different problem domains [145, 146]. This work employs genetic algorithms and learning classifier systems based upon their specific features described in chapter 1 section 1.2.

### 2.3.1 Genetic Algorithms

Genetic algorithms (GAs) are search based heuristic methods based on the principle of natural selection [51]. GAs encode the decision variables of a search task into finite-length strings of certain cardinality [120]. These strings are composed of alphabets. The strings being candidate solutions to the search problem are termed as chromosomes, the alphabets of the string are called genes and the value of genes are termed as alleles. For instance, in the well-known travelling salesman problem, a chromosome represents a route and a city may be represented by a gene. Contrary to traditional optimisation schemes, GAs work with coding of parameters, rather than parameters themselves [47].

To evolve good solutions, a fitness measure is utilised by the GA to guide the evolution of solutions. In contrast to traditional search methods, GAs rely on a population of candidate solutions. User defined population size is one of the most important factors affecting the performance and scalability of GAs. For example, small population sizes may lead to premature convergence leading to substandard solutions, while large populations might lead to wastage of computational resources and time [120]. GAs evolve solutions to the search problem using these steps:

1. *Initialization.* The initial population is usually generated randomly across the search space but incorporating domain-specific knowledge or some other information is also permitted.
2. *Evaluation.* After population initialization, the fitness of all candidate solutions is evaluated.
3. *Selection.* Selection imposes survival-of-the-fittest strategy on candidate solutions and allocates offspring of solutions with higher fitness value. The main idea of selection is to prefer better solutions over worse ones. Many selection procedures are proposed in literature including roulette-wheel selection, stochastic universal selection, ranking selection and tournament selection [47]. Tournament selection, which is used in this study

is explained below.

In tournament selection [47],  $s$  solutions are chosen at random and entered into a tournament. The fittest individual wins the tournament and selected as parent. Most widely used value of  $s$  is 2. Using this selection procedure a total of  $n$  tournaments are required to choose  $n$  individuals.

4. *Recombination or crossover.* Recombination or crossover combines two or more parental solutions to create new, possibly better solutions. Performance of methods proposed in literature is dependent on design of recombination mechanism. The offspring produced as a result will not resemble any particular parent but will combine parental traits [48]. A brief introduction of the special crossover operator (Laplace crossover) used in this study is presented here.

The Laplace crossover operator is self-adaptive and places the offspring proportional to the ranking of parents. If parents are close in the solution space, then the offspring have a high probability of being clustered and if parents are far the offspring are expected to be scattered [34]. This strategy favours such problems when there is not enough *a priori* knowledge about the relation of decision variables, therefore placing the offspring with respect to parents (which are the best individuals in the population) gives better chances of maintaining the schema.

5. *Mutation.* Unlike crossover, mutation locally modifies a solution. Most of the proposed methods usually perform a random walk in the vicinity of the parent solution. A brief introduction of the special mutation operator (power mutation) used in this study is included here.

The power mutation operator proposed in [33] is parent-centric and generates the offspring in close neighbourhood to the parent solution using a random variable that follows the power distribution.

6. *Replacement.* The new population generated by selection, crossover and mutation replaces the original population.

7. Steps 2-6 are repeated until a terminating condition is met.

GA's have been widely adapted for solving unconstrained and con-

strained optimisation problems. GA's have also proven to be useful for solving mixed integer programming problems where a few or all decision variables are integers [33]. The decision variables may also have bound constraints. In this thesis, the joint learning of feature related parameters and the feature importance is formulated as an optimisation problem with integer and bound constraints on decision variables. A GA based framework is proposed as a solution (please see chapter 3 and chapter 5) due to its promising performance in similar optimisation problems related to other engineering domains [33]. Additionally a multiple GA based formulation is introduced in this work to learn multiple feature importance rules (see chapter 6).

### 2.3.2 Learning Classifier Systems

Traditionally, a Learning Classifier System represents a rule-based agent that incorporates evolutionary computing and machine learning to solve a given task by interacting with a previously unknown environment. After observing the current state of the environment, the agent performs an action and the environment provides a reward. The generalization property in LCS allows a single rule to cover more than one state provided that the action-reward mapping is similar.

XCS [137] is a formulation of LCS that uses accuracy-based fitness to learn the problem by forming a complete mapping of states and actions to rewards. Accuracy based XCS system computes the fitness measure by utilising the accuracy of rules' prediction of expected payoff to create a map of the whole problem space rather than the traditional search for only high payoff rules. In XCS, the learning agent evolves a population  $[P]$  of classifiers, where each classifier consists of a rule and a set of associated parameters estimating the quality of the rule. Each rule is of the form 'if *condition* then *action*'. Traditionally, the condition is represented by a fixed length bit-string defined over the ternary alphabet  $\{0, 1, \#\}$  where '#' is



the ‘don’t care’ symbol which can be either 0 or 1; and the action is represented by a numeric constant. Each classifier has three main parameters: 1) fitness  $f$ , which provides a measure of classifier’s usefulness; 2) prediction error  $\epsilon$ , error between the predicted payoff and the actual received reward; 3) prediction  $p$ , which provides an estimate of the expected payoff from the environment, if the classifier’s action is executed. Each classifier also keeps track of the number of times it has been updated using the experience parameter  $exp$  and also the number of copies of a classifier in the population, denoted by  $n$ .

The system has two modes of operation namely explore (training) and exploit (application). In the explore mode, the system performs the following operations:

1. observes the current state  $s \in \mathcal{S}$  from the environment, where  $\mathcal{S}$  is the set of all possible states.
2. creates a match set  $\mathcal{M}$  by selecting the classifiers that match the current state  $s$ .
3. if action  $a_i$  is not present in  $\mathcal{M}$ , a random classifier is generated which matches state  $s$  and also advocates  $a_i$ .
4. a prediction array  $P(a_i)$  is generated that estimates the payoff given an action  $a_i$  is executed.
5. action  $a$  is selected for exploration and all the classifiers in  $\mathcal{M}$  are selected and placed in the action set  $\mathcal{A}$ .
6. after performing the action  $a$ , the reward  $r$  received from the environment is used to update the parameters of all classifiers in  $\mathcal{A}$ .
7. When deemed necessary, new classifiers are introduced into the population by rule discovery (usually genetic algorithm) in the action set  $\mathcal{A}$ .

In the explore mode, the agent performs the action with the best predicted payoff, instead of searching for new information.

Various richer encoding schemes have been investigated to represent high level knowledge in LCS in an attempt to obtain compact classifier rules [31, 62, 63], to solve overlapping problems [59, 64], to approximate functions [19, 139], to develop useful feature extractors [6], to tackle problems involving large number of discrete actions [96, 84], to compute continuous actions [140, 129, 61], and to identify and process building blocks of knowledge [20, 60].

The following XCS-based classifier system, i.e. XCSCA (XCS with Computed Action) [84], is directly related to the work to be presented here.

### **XCS with Computed Action**

XCSCA [84] is a supervised learning system, which computes action mappings using the input message  $x$  and a weight vector  $w$ . XCSCA borrows the idea of supervised learning from UCS [11]<sup>5</sup>, and the idea of action mappings from XCSF [139]<sup>6</sup>. In XCSCA, classifiers have no prediction since, as in UCS, there is no incoming reward. This is because XCSCA does not produce a complete action map, rather it only evolves the correct output function.

The classifiers have similar parameters as in XCS [137] with an additional parameter vector  $w$ , which is used to compute a discrete classifier action. XCSCA works similar to UCS for building the match set. In covering, classifier conditions and parameters are initialized as in XCS, while  $w$  is in with zero values. XCSCA employs two modes of operation: *learning*

---

<sup>5</sup>UCS is the extension of XCS to supervised learning for multi-class problems and problems with unbalanced classes, proposed developed by Bernado et al. [11]. It evolves a more efficient action map and replaces the reinforcement learning scheme of XCS with a supervised learning scheme.

<sup>6</sup>XCSF extends XCS by introducing computed prediction for learning approximations of functions, developed by Wilson [139]. It adds a weight vector to the classifiers and replaces the scalar prediction with a computed prediction.

*mode* and *testing mode*. During learning similar explore and exploit phases are executed as in XCS. During test mode, XCSCA computes the discrete action  $a$  for each classifier in  $\mathcal{M}$ . For each action  $a$  for the classifiers in  $\mathcal{M}$ , XCSCA computes the classification accuracy of  $a$  for the input  $x$ . Finally, the action with the highest classification accuracy is selected.

The framework introduced in this work utilises XCSCA to learn multiple feature importance rules by grouping input images into niches using a novel input matching scheme (please see chapter 7).

## 2.4 Chapter Summary

The majority of the saliency detection techniques introduced in literature for salient object detection have limited generalisability as they employ ad hoc techniques for feature normalisation, weighting and integration [72]. Previous learning based methods [13, 72] approach the problem of feature importance learning using traditional classifiers. However, the problem of learning feature importance and related parameters is an optimisation problem in its essence, where the bounds on the parameters are the constraints to be taken into account. Genetic Algorithms are highly competitive constrained optimisation methods that are best suited for searching the optimal solution using the whole population. Due to their inherent formulation for constrained optimisation and effective search capabilities, they can be readily employed for joint learning of feature importance and related parameters.

Despite of the limited efforts to increase generalisation of the saliency detection methods [13, 68, 72], no investigation has been reported for learning multiple feature importance rules each suited to a type of image to increase generalisation to unseen image data. Multiple genetic algorithms are to be employed for the first time to learn multiple feature combination schemes and a new grouping scheme is introduced to group similar images (chapter 6). Moreover, the inherent property of XCS to divide the

search space into niches is to be employed to group images and learn multiple feature combination rules (chapter 7).

A few methods have been introduced for selecting the best features during fusion and discarding the irrelevant features [49, 27]. However, they are limited in terms of judging true quality of features for salient object detection hence select false features during fusion. A new dynamic feature selection method based on novel feature quality measurement cues is introduced in the work presented in this thesis (chapter 8).

Additionally as most methods are region-based, non-uniform highlighting of the salient object is a common artifact of saliency responses [26]. A matting components based method inspired by spectral clustering approaches is introduced in this work to overcome the limitation of region based methods (chapter 4).

## Chapter 3

# Optimizing System Parameters of the Visual Attention Model for Human Fixation Prediction

### 3.1 Introduction

The previous chapter presented a detailed review of the feature integration theory and the traditional model of visual attention. It also covered design and details of features that have been proposed by various techniques for different visual saliency applications. An introduction was presented about the various tunable design parameters of the traditional model of attention and the choice of suitable values for these parameters in accordance with prior work. Moreover, techniques for dealing with the non-trivial problems of feature conditioning and feature combination were discussed.

The traditional computational model of visual attention [66] has tunable parameters as depicted in Figure 3.1. The filter parameters inside the dashed box are employed by the orientation feature channel to compute the orientation feature maps. The set of weights  $\{wC_v, wI_v, wO_v\}$  are used to assign importance to the colour, intensity and orientation features re-

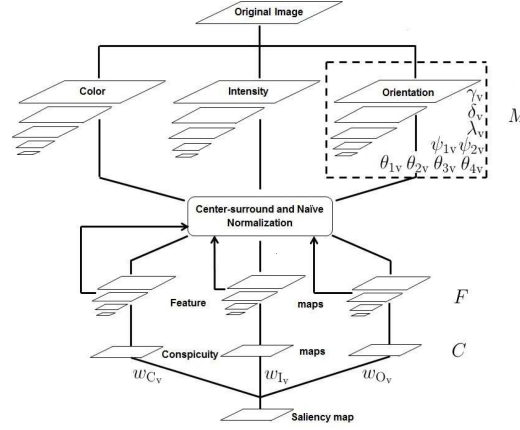


Figure 3.1: Computational model of visual attention [66]. Tunable parameters are shown.  $M$ ,  $F$  and  $C$  represent the raw maps, feature maps and the conspicuity maps, respectively.

spectively. The tunable parameters at the feature computation stage and the feature importance parameters at the combination stage are jointly defined as system parameters.

Previous works such as the work of Itti et al. [66] and the work of Walther and Koch [134] employed human encoded values for the tunable filter parameters and assigned equal importance to all the features. Due to human chosen parameters these methods are subject to low generalisation as these parameters require tuning in an *ad hoc* fashion to work well. To the best of author's knowledge, there has been no prior effort to jointly learn the tunable parameters at the feature computation stage and the feature importance weights at the combination stage.

It is hypothesised that the joint learning of the feature computation parameters and the feature importance parameters can improve the saliency detection performance of the traditional model of visual attention through optimization of a task-specific objective function<sup>1</sup>. The basis of this hy-

<sup>1</sup>Task-specific objective function here refers to an objective function designed especially for a specific saliency detection task, such as eye fixation prediction or salient object

pothesis is that suitable values for feature related parameters improve the saliency detection performance of the features, hence the feature importance needs to be reinforced accordingly. The problem of searching for the best feature computation parameters along with appropriate feature importance weights can be naturally formulated as an optimisation problem. Given an appropriate training set, the best suited parameters and weights are expected to increase the generalisation of the system on unseen images. Previous works have formulated the problem of learning features weights as a classification problem by employing support vector machines SVMs [72] and AdaBoost [13]. However, these classification methods are not adaptable for the above-mentioned problem due as it involves additional parametric search along with weight learning. Thereby, a technique is needed that has the ability to effectively encode parameters with different modalities and is able to competitively search the feasible solution space that is defined by the bounds on the parameters. Additionally, the optimisation method must have the ability to incorporate objective functions that directly encode important performance measures related to saliency detection such as area under the precision recall curve (AUCPR). instead of objective functions that incorporate indirect performance measures such as class separability.

Amongst other optimisation techniques, Genetic Algorithms (GA) are inherently a good choice, as they have the ability to deal with the large search space constituted by the system parameters, are competitive in large populations and they also provide the flexibility to encode unique tasks as their objective function.

It is worth mentioning that GA is not necessarily the only feasible solution to this problem and other optimisation methods may prove to be equally competitive. The main contribution of this work is the introduction of joint learning framework in the traditional computational model of visual attention. Therefore no specific investigation is conducted on find-

---

detection.

ing the best optimisation method for the problem at hand.

Following the above discussion, the goal of this chapter is to develop a GA based optimization framework that is able to extend the traditional model of attention by learning system parameters for the task of human fixation prediction. To achieve this goal, a new GA based framework is developed by designing an objective function that searches for suitable values for the system parameters by maximizing the agreement between the predicted saliency map and the human fixation ground truth. The designed GA framework is deployed to search for optimal<sup>2</sup> parameters for a baseline saliency method. Specifically, the first objective of this chapter is to compare the performance of an optimised feature set (that employs learned parameters) with the unoptimised feature set (that uses human encoded parameters) for human fixation prediction task. The second objective is to compare the performance of the optimised visual attention method against the unoptimised baseline method for the task of fixation prediction.

### 3.1.1 Chapter Organisation

The remainder of the chapter is organised as follows: Section 3.2 details the proposed GA based algorithm and details the implementation of the visual attention method for the proposed method. Section 3.3 presents the design of experiments. Section 3.4 presents and discusses the obtained results, before the final section summarises the findings of this chapter.

---

<sup>2</sup>The term optimal is used in this thesis to describe the best possible solution given constraints on computational time and power available to search. This may not be the global optimum. The methods employed in this work do not ensure global optimum on the search space.



## 3.2 The Proposed Genetic Algorithm Optimised Visual Saliency Method (GAOVSM)

In this section, the details of the visual attention method implemented in this study are first presented, including a description of the tunable parameters that need to be searched. Next, the details of the new genetic algorithm designed to learn optimal solutions for these parameters, termed as Genetic Algorithm optimised visual saliency method (GAOVSM), are presented.

### 3.2.1 The Visual Saliency Method

The visual saliency method implemented in this study follows the standard Itti-Koch saliency model [66]. The efficacy of the model has been extensively tested in prior works as demonstrated by its recent application in eye fixation prediction, such as guiding fixations [14] and gaze learning [110]. Other recent practical applications include keypoint detection [40] and assistive technologies for the visually impaired [125]. Specifically, the simple implementation from the work of Walther and Koch [134] is adapted in this work. This implementation was chosen, because a primitive implementation is most suitable for this work as compared with other implementations [66, 50] that include several add-ons and complex functionality.

The implementations details of the method are the same as discussed in the background chapter. There are multiple tunable parameters of the orientation channel as depicted inside the dashed box in Figure 3.1. These features include aspect ratio ( $\gamma_v$ ), standard deviation ( $\delta_v$ ), wavelength ( $\lambda_v$ ), orientations ( $\theta_{1v}, \theta_{2v}, \theta_{3v}, \theta_{4v}$ ) and the phases ( $\psi_{1v}, \psi_{2v}$ ). Suitable values for these parameters are searched during the optimisation process at the training stage, while the test stage employs the learned values.

The complete parameter set  $\phi$  to be learned during the optimisation

stage is given as follows:

$$\phi = \{\gamma_v, \delta_v, \lambda_v, \theta_{1v}, \theta_{2v}, \theta_{3v}, \theta_{4v}, \psi_{1v}, \psi_{2v}, w_{C_v}, w_{I_v}, w_{O_v}\}, \quad (3.1)$$

where the subscript  $v$  is employed to depict that these parameters are to be searched for optimal values.

### 3.2.2 Genetic Algorithm

A new implementation of a genetic algorithm (GA) is introduced in this work to learn optimal solutions for the parameter set given by equation (3.1) (which also defines the dimensionality of the search space). A real-coded implementation for the GA is employed to avoid the Hamming Cliff problem<sup>3</sup> and the processing time overhead of the binary-coded GAs [33]. The parameters constitute the decision variables and are encoded as integers in the real coded representation. This is achieved by enforcing integer restrictions on every individual after crossover and mutation operations. As sufficient knowledge of the solution is not available *a priori*, a random initial population is created with the parameter values drawn from a uniform distribution. It is ensured that the population is bounded by the constraints on the variables. As the objective function (see section 3.2.2) returns a fitness in the range [0,1], rank scaling is employed to facilitate the selection process by enhancing the discriminability between the solutions. Elitism is utilised to ensure good solutions are present to guide the search throughout the evolutionary process. Laplace crossover and power mutation functions are employed in order to drive the search in a parent-centric fashion (for details please see sections 3.2.2 and 3.2.2). The number of generations and a tolerance value for average change in fitness value

---

<sup>3</sup>The hamming distance between two consecutive binary strings is the number of bits that must be converted from one to another in order to make the two strings equal. In binary representation for GAs, large hamming distances between two consecutive binary strings are referred as “hamming cliffs” as the GA must alter numerous bits simultaneously to obtain a transition between such strings.

over a defined number of generations constitute the stopping criteria. Details of the design choices for the parameters and GA methods are now presented.

### Parameter Encoding

The chromosomes are encoded using a real coded representation. In common with other approaches, rounding off of real variables is used to handle integer constraints on decision variables. The ranges of the decision variables are based on the widely employed ranges for these parameters and are inspired by past studies [57, 70, 132].

The aspect ratio of the filter  $\gamma_v$  determines the ellipticity of the receptive field and is defined as:

$$\gamma_v \in [0.5, 1]. \quad (3.2)$$

The standard deviation of the Gaussian  $\delta_v$  and the wavelength  $\lambda_v$  determine the bandwidth  $B$  of the filter. The bandwidth of the filter determines how the Gabor filters cover the visual field in the frequency space. In this study, the range of standard deviation and wavelength is set to limit the bandwidth  $B$  of the filter from 0.2 to 4.0 octaves according to prior work [132]

$$\left. \begin{array}{l} \delta_v \in [\delta_{v,\min}, \delta_{v,\max}] \\ \lambda_v \in [\lambda_{v,\min}, \lambda_{v,\max}] \end{array} \right\} \mid 0.2 < \log_2 \frac{\frac{\sigma}{\lambda} + \frac{1}{\pi} \sqrt{\frac{\ln 2}{2}}}{\frac{\sigma}{\lambda} - \frac{1}{\pi} \sqrt{\frac{\ln 2}{2}}} < 4.0. \quad (3.3)$$

The range of the bandwidth is decided according to physiological studies on Macaque visual cortex, which can be generalised to human vision as claimed by Valois et al. [132]. The lower and upper ranges of the parameters  $\delta_v$  and  $\lambda_v$  are set as to satisfy equation 3.3.

Previous approaches generally use four fixed orientation scales only e.g.  $0^\circ, 45^\circ, 90^\circ, 135^\circ$  to construct the orientation pyramids. Based on some previous approaches such as [112, 141], which make use of various discrete orientations like  $0^\circ, 4^\circ, 8^\circ, 16^\circ$  and  $0^\circ, 36^\circ, 72^\circ, 108^\circ, 144^\circ$  respectively; the orientations  $\theta_{1v}, \theta_{2v}, \theta_{3v}, \theta_{4v}$  are allowed to vary from  $0^\circ$  to  $180^\circ$ .

Based on past works [57], the range for the phases  $\psi_{1v}, \psi_{2v}$  of the Gabor filters is set as following:

$$\psi_{1v}, \psi_{2v} \in [-180 \cdots 180] . \quad (3.4)$$

Finally, the weights for the intensity, colour and orientation features are bounded as follows and encoded as real values in the GA;

$$w_{I_v}, w_{C_v}, w_{O_v} \in [0, 1] . \quad (3.5)$$

### Objective Function

To quantify the fitness of a computed saliency map, its objective evaluation in terms of its agreement with target fixation locations is needed. This is achieved by employing the saliency map as a classifier (via thresholding). All the points which are above a particular threshold value are classified as targets (fixations) and all other as background (no fixations). The fraction of real targets classified as targets for a particular threshold value define the true positive rate, while the fraction of background points classified as targets give the false positive rate. A receiver operating characteristic (ROC) curve is generated by utilizing multiple thresholds as operating points. The obtained area under ROC curve  $AUC_{\text{Computed}}$  is then utilised to compute the fitness value as:

$$F(\phi) = \arg \min_{\phi} (AUC_{\text{Ideal}} - AUC_{\text{Computed}}), \quad (3.6)$$

where  $AUC_{\text{Ideal}}$  is 1.

### Selection

Binary tournament selection [47] is employed to select the individuals for subsequent generations. In our experiments a selection pressure provided by a tournament size of two puts sufficient selection pressure on the most fit individuals. Choosing the better solutions as parents and parent-centric crossover and mutation operators helps in maintaining the best schema.

### Crossover

Due to lack of *a priori* information about the relationship of decision variables, a parent-centric crossover operator as proposed in [33] is employed. The Laplace crossover operator is self-adaptive and places the offspring proportional to the position of parents. If parents are close, then the offspring have a high probability of being clustered and if parents are very different then the offspring are also expected to be scattered [33]. This strategy favours the problem at hand as there is not enough *a priori* knowledge about the relation of decision variables, therefore placing the offspring with respect to parents (which are the best individuals in the population) gives better chances of converging to the optimum.

Using Laplace crossover [32], parents  $\hat{u}^{(1)} = (\hat{u}_1^{(1)}, \hat{u}_2^{(1)}, \dots, \hat{u}_n^{(1)})$  and  $\hat{u}^{(2)} = (\hat{u}_1^{(2)}, \hat{u}_2^{(2)}, \dots, \hat{u}_n^{(2)})$  are used to generate two offspring  $\hat{v}^{(1)} = (\hat{v}_1^{(1)}, \hat{v}_2^{(1)}, \dots, \hat{v}_n^{(1)})$  and  $\hat{v}^{(2)} = (\hat{v}_1^{(2)}, \hat{v}_2^{(2)}, \dots, \hat{v}_n^{(2)})$ . A random number  $\Gamma$  is generated based on a uniformly distributed random number  $\xi \in [0, 1]$ .  $\Gamma$  follows the Laplace distribution given as inverse function of the Laplace distribution:

$$\Gamma = \begin{cases} a - b \log_e(\xi), & \xi \leq \frac{1}{2}, \\ a + b \log_e(\xi), & \xi > \frac{1}{2}, \end{cases} \quad (3.7)$$

where  $a$  and  $b$  are the location and scale parameters of the Laplace distribution. The offspring are then computed as follows:

$$\hat{u}_i^{(1)} = \hat{v}_i^{(1)} + \Gamma \left| \hat{v}_i^{(1)} - \hat{v}_i^{(2)} \right|, \quad (3.8)$$

$$\hat{u}_i^{(2)} = \hat{v}_i^{(2)} + \Gamma \left| \hat{v}_i^{(1)} - \hat{v}_i^{(2)} \right|. \quad (3.9)$$

### Mutation

On similar grounds, the mutation process needs to be guided by the position of the parent solution. Therefore, the power mutation operator proposed in [35] is employed. This approach is also parent-centric and gen-

erates the offspring in close proximity to the parent solution. Given uniformly distributed random variables  $\psi, \varphi \in [0, 1]$  and a random variable  $\Xi$ , which follows the power distribution, the mutated solution is computed as follows [32]:

$$\hat{u} = \begin{cases} \hat{v} - \Xi(\hat{v} - v^l), & \psi < \varphi, \\ \hat{v} + \Xi(v^u - \hat{v}), & \psi \geq \varphi, \end{cases} \quad (3.10)$$

where  $\psi = \frac{\hat{v} - v^l}{v^u - v^l}$  and  $v^l, v^u$  are the lower and upper bounds of the decision variable.

### 3.3 Design of Experiments

#### 3.3.1 Data Set

The work described in this chapter uses human eye-tracking data from three challenging data sets of the ImgSal database [90]. The ImgSal database comprises of 235, 480  $\times$  640 pixel images collected from Google and other literature sources, which are divided into six categories. The datasets contain fixation data from 19 naïve subjects who were just asked to observe the images with no specific instructions. Unlike other databases, ImgSal takes into account the difficulty in analysing images by classifying images into the following categories: images with both large and small salient regions, images with cluttered background and images with repeated distractors. A good saliency method should perform well in all the above conditions [90].

For the aforementioned reason, the three most challenging categories form the saliency database termed as Cat1, Cat2 and Cat3, respectively, are employed in this work. Cat1 contains 50 images with large salient regions. Cat2 contains 15 images with cluttered background. Cat3 contains 15 images with both large and small salient regions.

### 3.3.2 Ground Truth Data

Ground-truth fixation maps are also provided by ImgSal [90]. The fixation maps are of the same resolution (480x680) as the original images. The fixation maps contain processed fixation data, where each location represents a number, which is the count of fixations from all participants for that particular location. For qualitative comparison, the fixation maps were blurred using the standard Gaussian filtering procedure and heat maps were created to represent the human eye fixation data. The standard parameters from the literature were used for Gaussian filtering [72].

### 3.3.3 Performance Measures

In this study we have used three performance metrics namely area under the Receiver Operating Characteristics (ROC) curve (AUC), a similarity measure  $S$  [126] and Normalised Scanpath Saliency (NSS) [114] as described below. As AUC is involved in the fitness computation for the proposed approach, the two independent performance metrics namely  $S$  and NSS are also employed to effectively evaluate performance without any bias. These measures provide an extra dimension to the AUC evaluation measure by measuring how similar are the distributions of the predicted saliency and ground truth, and a measure of agreement between the human scanpath and the predicted map.

The similarity measure ( $S$ ) is adapted from a recent study [126] and is used to evaluate the similarity of two distributions. Both the distributions are normalised so that they sum to one. Afterwards  $S$  is computed by summing the minimum values at each point in the distributions [126]. If  $S$  is evaluated to be 1, it implies that the distributions are the same, while an  $S$  score of 0 indicates that the distributions are entirely different [126].

NSS gives a measure of agreement between human fixation locations and method predictions [114]. A score greater than 0 means a higher correspondence, 0 means no correspondence and a score less than 0 shows anti-

correspondence between method predicted output and human scanpaths. It is calculated by normalizing the saliency map of the method to have a zero mean and unit standard deviation. Afterwards, a human subject's scanpath is used to access the corresponding normalised saliency values at each point. The mean of the extracted normalised salience values gives the NSS score [114].

### 3.3.4 Experimental Settings for the Genetic Algorithm

The population size for the GA was set to 100 and the GA was run for 20 generations for each experiment. Increasing population size beyond 100 individuals resulted in slowing down the convergence rate without any significant increase in performance. Similarly running the evolutionary process for more than 20 generations also did not improve the performance. The crossover rate was set to 0.8 and the mutation rate was set to 0.01 after empirical testing.

The first experiment was performed on Cat1 images with large salient regions. The data set of 50 images was divided into 35 training and 15 testing images. 30 runs were repeated with different random seeds and the GA was trained to search for the optimal solution that enhances the performance.

The second experiment was performed using Cat2 images that have a cluttered background. The data set contains 15 images in total. It was divided into 11 training images and four testing images. All the parameters of the GA were kept the same. Again, the average of 30 runs of the GA was evaluated to search for optimal solutions for this dataset.

The third experiment was performed on Cat3 images with both large and small salient regions. The Cat3 data set consists of 15 images in total divided into 11 training and 4 testing images as Cat2. As above, the same parameters were used in this experiment.

The datasets used for validation are small but are used in standard



practice for the evaluation of saliency methods [90]. The training and testing sets chosen in this work are also small, however have been employed for evaluation in similar settings by prior works [92].

### 3.3.5 Selected Visual Attention Methods for Comparison

The performance of the proposed GAOVSM approach is compared with eight state-of-the-art methods including RARE [117], AIM [108], Hou and Zhang (HZ) [54], Seo and Peyman (SP) [121, 122], Torralba [128], SUN [152], Achanta [3] and STB [134]. These methods are chosen because of multiple reasons 1) almost all of them are heuristic and do not involve learning and optimisation in contrast to the proposed GAOVSM. It is noted that comparison of a learning based method with heuristic methods is against the norm, it is performed here to assess whether the incorporation of learning improves the performance over traditional heuristic methods. 2) All of the compared methods were primarily developed for the task of fixation prediction; 3) Saliency maps for these models are either available online or could be generated easily from the author provided implementations.

## 3.4 Results and Discussion

In this section, firstly the performance of the orientation feature when computed with human encoded parameters is compared with its performance when computed using the optimisation based learned parameters in section 3.4.1. Next, an analysis of the evolved solutions is presented to reveal an insight into the improvements of the proposed method. Finally, quantitative and qualitative comparison of the proposed approach with heuristic state-of-the-art methods is reported in section 3.4.3.

### 3.4.1 Performance Comparison of Orientation Feature, Learned Versus Human Encoded Parameters

The performance of orientation feature computed using learned parameters as compared with its performance using human encoded parameters is shown in Figure 3.2. Figure 3.2 (a) shows the average ROC curves on images from all three categories for optimised and human encoded parameters. The orientation feature with optimised parameters outperforms the feature computed with human encoded parameters according to our initial hypothesis. The yellow region shows the gain in performance achieved by the orientation feature with optimised parameters over its counterpart. Figure 3.2 (b) shows the performance comparison in terms of additional evaluation measures such as average AUC, similarity and NSS for individual image categories. The average results obtained for the unoptimised orientation feature are 0.6118, 0.2184 and 0.7087 in terms of AUC, similarity and NSS, respectively. The orientation feature with optimised parameters improves upon the unoptimised feature with average measures of  $0.7499 \pm 0.0130$ ,  $0.3865 \pm 0.0068$  and  $1.0660 \pm 0.1433$ . The figures after the  $\pm$  sign represent the 99% confidence interval on the mean.

### 3.4.2 Analysis of Evolved Solutions

Representative parameter sets  $\phi$  produced by the GA after optimization through successive generations for the three image categories are shown in Table 3.1. For all categories, the optimised standard deviation  $\delta^*$  and wavelength  $\lambda^*$  parameters were selected by the proposed GAOVSM approach to provide a low bandwidth. This choice of  $\delta^*$  and  $\lambda^*$  corresponds to narrow-bandwidth filters for object background discrimination in cluttered scenes. This result is in agreement with a previous study, which used narrow-bandwidth filters for discriminating objects from background in cluttered scenes [17]. The optimised aspect ratio  $\gamma$  is found to be relatively high for all categories, representing a receptive field tuned to elon-

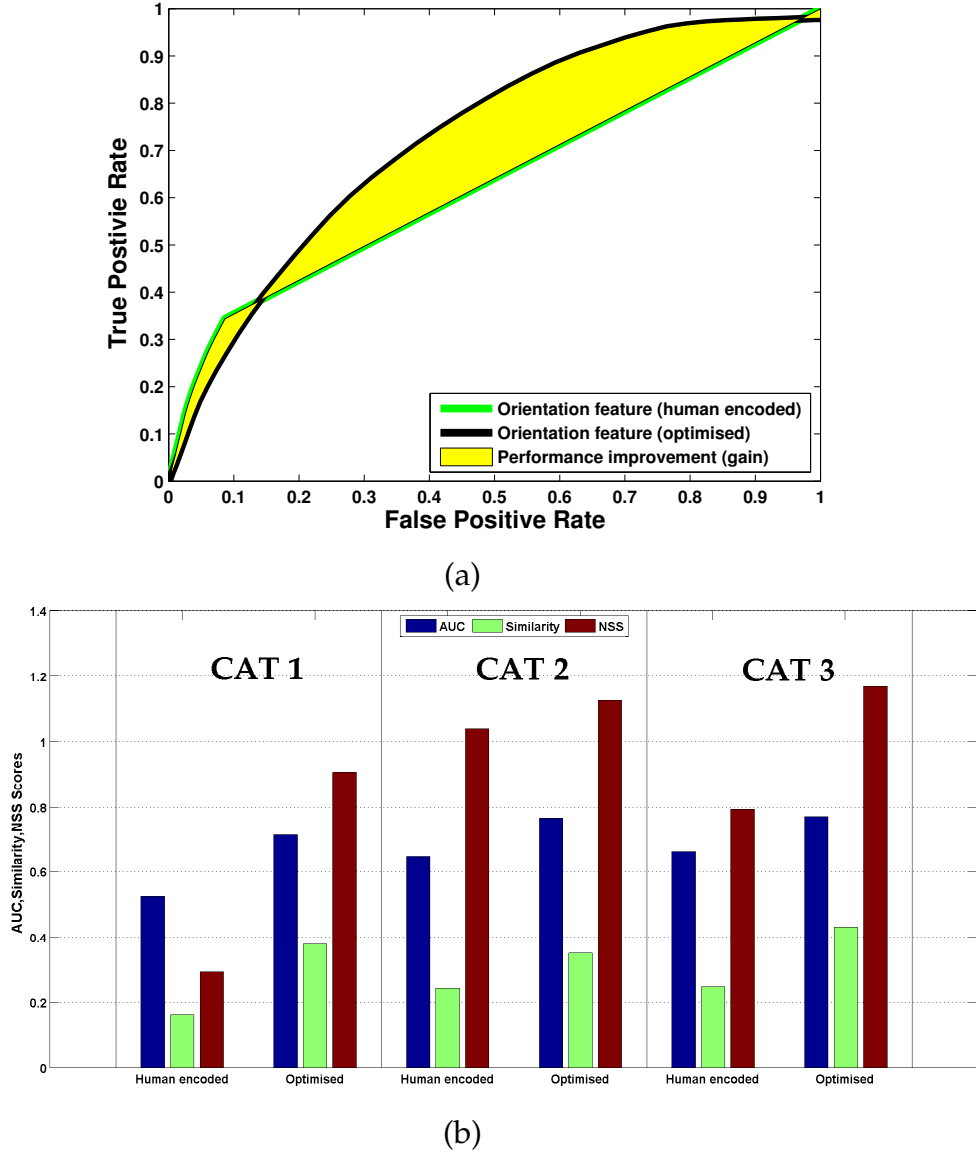


Figure 3.2: (a) Average ROC curves for the orientation feature (human encoded (green) vs optimised parameters (black)) on all three categories of ImgSal. (b) AUC, similarity and NSS scores for the orientation feature (human encoded vs optimised parameters) for all three categories of ImgSal.

gated contours of objects.

Table 3.1: Representative optimised parameters  $\phi$  evolved by the GA for all categories.

Sr.No.	Parameter Name	Optimised Values Cat1	Optimised Values Cat2	Optimised Values Cat3
1	Aspect Ratio $\gamma^*$	0.65	0.95	0.95
2	Standard Deviation $\delta^*$	7.93	6.95	7.57
3	Wavelength $\lambda^*$	6	6	6
4	Orientation1 $\theta_1^*$	$3^\circ$	$40^\circ$	$42^\circ$
5	Orientation2 $\theta_2^*$	$77^\circ$	$87^\circ$	$84^\circ$
6	Orientation3 $\theta_3^*$	$92^\circ$	$104^\circ$	$100^\circ$
7	Orientation4 $\theta_4^*$	$135^\circ$	$169^\circ$	$138^\circ$
8	Phase1 $\psi_1^*$	$-150^\circ$	$18^\circ$	$-173^\circ$
9	Phase2 $\psi_2^*$	$-169^\circ$	$109^\circ$	$-103^\circ$
10	Weight Colour $W_C^*$	0.27	0.43	0.33
11	Weight Intensity $W_I^*$	0.10	0.05	0.24
12	Weight Orientation $W_O^*$	0.88	0.70	0.94

Optimised orientations  $\theta_2^*$  and  $\theta_3^*$  are found to be favouring the vertical orientation. This result corresponds to the oblique effect [89]. According to the oblique effect, there is a bias in the human visual cortex for horizontal and vertical orientations as compared with the oblique ones. Orientations  $\theta_1^*$  and  $\theta_4^*$  are found to be consistent with the human encoded values in general.

For Cat2 images, the optimised phases  $\psi_1^*$  and  $\psi_2^*$  are found to be symmetric. This is a reasonable result as symmetric filters have been applied in the past for object detection in presence of clutter [116]. For Cat1 and Cat3, the optimised phases were found to be mixtures of asymmetric filters. This is an inherent fit for Cat1 and Cat3 (containing images with different object sizes), as the mixtures of asymmetric filters are sensitive to

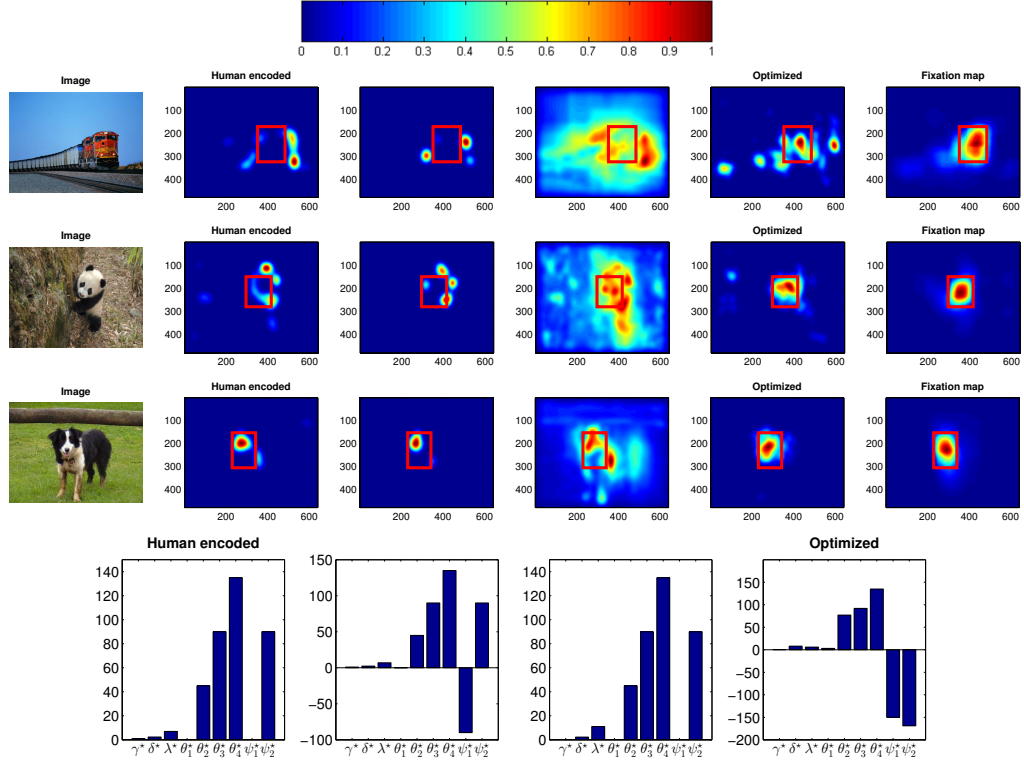


Figure 3.3: Effect of parameters on the orientation feature conspicuity maps for three representative images from Cat1. For each image four maps are computed using the parameter settings shown in the fourth row. The red boxes are bounded boxes for the high intensity regions in the ground truth fixation maps.

object sizes in images.

Finally the orientation feature is most heavily weighted as compared with the other features. This results confirms a previous study reported in [153]. It is worth noting that the orientation channel is also tuned in the proposed optimisation framework in contrast to the other channels.

To further analyse the effect of optimised parameters, Figure 3.3 shows the visual response of the orientation feature computed using four different parameter sets. The first parameter set contains human encoded

values. The second parameter set contains similar values to the human encoded values except for the phase values. The phases are set as anti-symmetric in contrast to symmetric values. The third set also includes similar values to human encoded values, where only the aspect ratio is decreased and bandwidth of the filters is increased. The final set contains the optimised parameter values.

It can be observed that visual response for the human encoded parameters tend to capture the edges of object contours (specifically for the first two images of Figure 3.3), while neglecting the conspicuous high contrast edges inside objects. This results in low agreement with the ground truth fixation maps. For the second set of parameters with antisymmetric phase values, the visual response is again similar to the human encoded one in that it also captures the object contours. Noticeable changes in the visual response can be observed, which can be attributed to the change in symmetry of the filters. For the third parameter set, the low aspect ratio and the high bandwidth results in a dispersed response. This response exhibits greater agreement with the ground truth as compared with the response of the first and second parameters sets. However this is achieved at the cost of including more background noise. In contrast to these parameter settings, it can be clearly observed that the visual response of the optimised feature set shows the best agreement with the ground truth. This was achieved by tuning the receptive field to be highly elliptical, with narrow bandwidth, preferring the vertical orientation (similar to the visual cortex) and employing asymmetric mixtures for phases.

### 3.4.3 Comparison of GAOVSM with State-of-the-art

The average ROC curves for all categories is shown in the first row of Figure 3.4. To provide a better picture of method rankings, the area under the ROC curves is shown in the second row of Figure 3.4 along with additional independent measures, i.e.  $S$  and NSS. The error bars on the bar

graphs show the 95% confidence level on the mean.

Firstly, we present the comparison of the proposed GAOVSM method with the baseline method of Walther and Koch [134], termed as STB. It is worth noting that the difference in performance of the proposed method in comparison to STB is substantial, despite having similar features and implementation. This result can be attributed to two key factors 1) the learned parameters of the proposed method are responsible for noteworthy gain in performance. This result is also supported by the performance gains achieved by the learned parameters regarding feature level evaluations; 2) the equally weighted feature maps in STB promote the chances of a poor feature map corrupting the final saliency response. Another factor which is noted in our experiments and results is that the feature maps of STB are subjected to excessive normalisation, resulting in a highly conservative final saliency map (for specific visual examples, please refer to Figure 3.5), which in turn causes STB to completely miss the highlighted regions in the ground truth map for certain images.

The proposed GAOVSM and the benchmark RARE method exhibit higher true positive rates for all the operating points on the ROC curve as compared with other methods, while showing similar performance to each other. For all the operating points in the ROC curve in terms of the thresholds, SP and HZ show similar true positive rates for corresponding false positive rates.

In terms of AUC, the proposed GAOVSM method shows better performance than all the other state-of-the-art methods, which is further supported by smaller deviation from the mean. This performance can be attributed to the newly designed objective function of the proposed method, which specifically optimises the AUC. In terms of  $S$ , the AIM method shows comparable performance as compared with GAOVSM, while GAOVSM outperforms all the other state-of-the-art methods. With reference to the NSS measure, RARE generally performs better than the other methods, however, GAOVSM shows more confidence and smaller deviation from

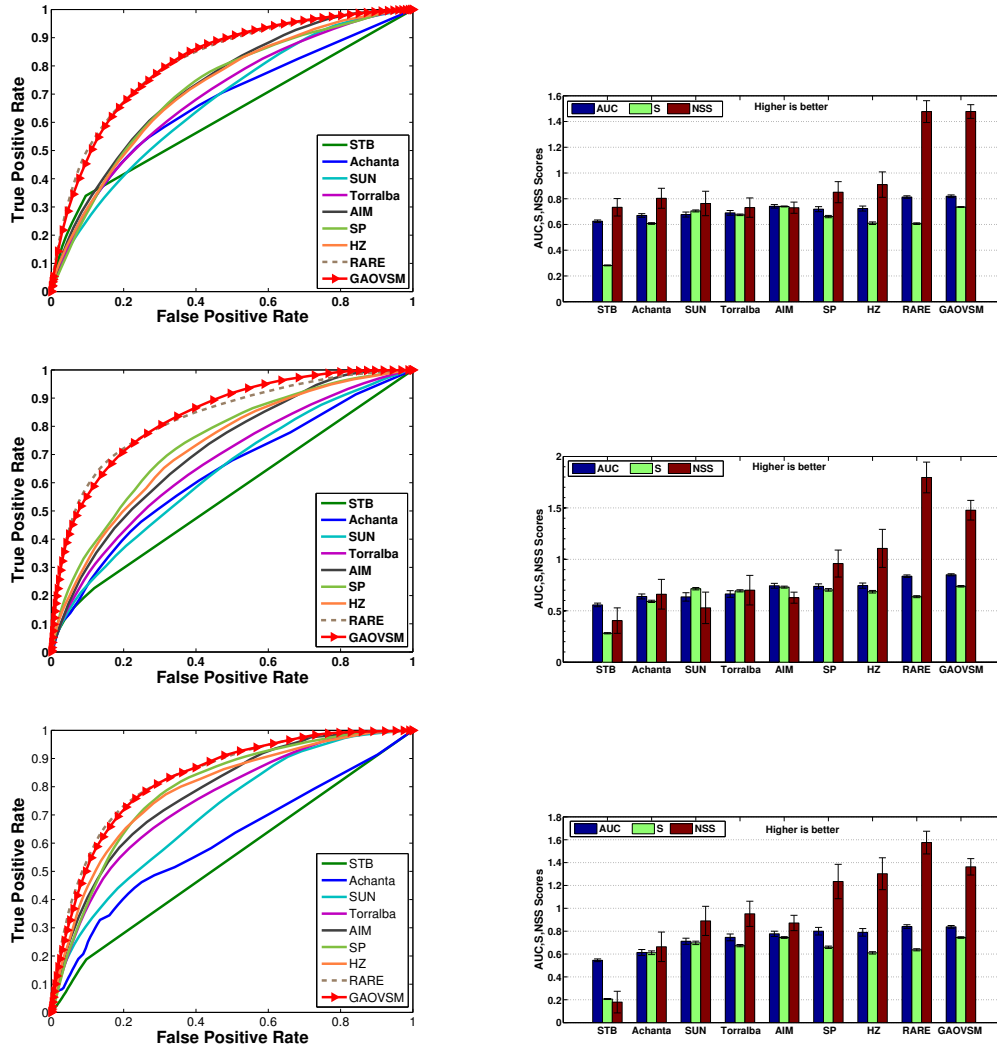


Figure 3.4: Performance comparison of methods on all image categories. The left column results show the average ROC curves for methods. The right column shows the results for average AUC,  $S$  and NSS measures for methods on all the image categories.



mean NSS. The reason for the high NSS score by the RARE method is that its internal normalisation already outputs maps with zero mean, therefore they are highly suited to the NSS measure, which performs similar normalisation as a pre-processing step before comparing it with the human's scanpath.

Examples of the visual output of the GAOVSM with the other state-of-the-art methods along with human fixation maps is shown in Fig. 3.5. A single image selected from each image category along with the corresponding visual response of attention methods is shown. For a method to perform well qualitatively, it must acknowledge the regions in the human fixation maps where the strongest activity is concentrated, shown by dark red colour in heatmaps. A few methods employ human encoded parameters during saliency computation. Specific examples of such methods are STB and SP. The highly sparse response of STB can be clearly observed, which may be attributed to the human encoded parameter that defines the amount of normalisation required. It can be observed from Figure 3.5 that the high saliency regions in the sparse STB maps do not overlap with the salient regions of the ground truth fixation maps.

SP employs human encoded values for its smoothing parameter and fall-off parameter in order to compute local steering kernels and self-resemblance, respectively. At the cost of false positives, SP manages to achieve considerable overlap with the ground truth human fixations as depicted by its quantitative performance in Figure 3.5.

The visual response of the AIM method exhibits maximum overlap with the ground truth fixation maps in areas of high saliency at the cost of including a considerable number of false positives. The SUN method efficiently matches the ground truth for images with easier background as can be seen from its response on the airplane image. However, for cluttered background images it exhibits an undesirable response by including background noise.

Although, the visual response of the RARE method highlights salient

object regions with high intensity blobs, it does not tend to capture high saliency regions from the eye fixation ground truth on all the representative images.

Although the visual output of GAOVSM is not very definitive in terms of defining objects as compared with some of the other methods, GAOVSM efficiently detects the main blobs of interest in the human eye-tracking data.

### 3.5 Chapter Summary

The goal of this chapter was to equip the traditional model of visual attention with the ability to learn the best suited values for its important parameters for improved performance. To achieve this goal, a new genetic algorithm was proposed to search for optimal values for the important parameters of the visual attention model. The objective function of the proposed GA was designed to maximise the agreement between the saliency output and the desired ground truth fixations by minimizing the difference between computed and ideal area under the ROC curve. The proposed method was able to show considerable feature and system level improvements as compared with its baseline counterparts. Additionally, the method exhibited comparable performance to several state-of-the-art saliency methods. The important findings of this chapter are:

It was shown that learning suitable values for important parameters of the feature computation process can yield substantial feature level performance improvements. At the feature level, quantitative improvements of 22.5%, 77% and 50% in terms of AUC,  $S$  and NSS measures were observed on average. The 99% confidence intervals on the mean suggests that the optimised features always substantially improve upon the performance of the baseline features for all runs. Additionally, it was shown that the optimised feature set produces maps that better match the desired ground truth at pixel level, as compared with the baseline method in order

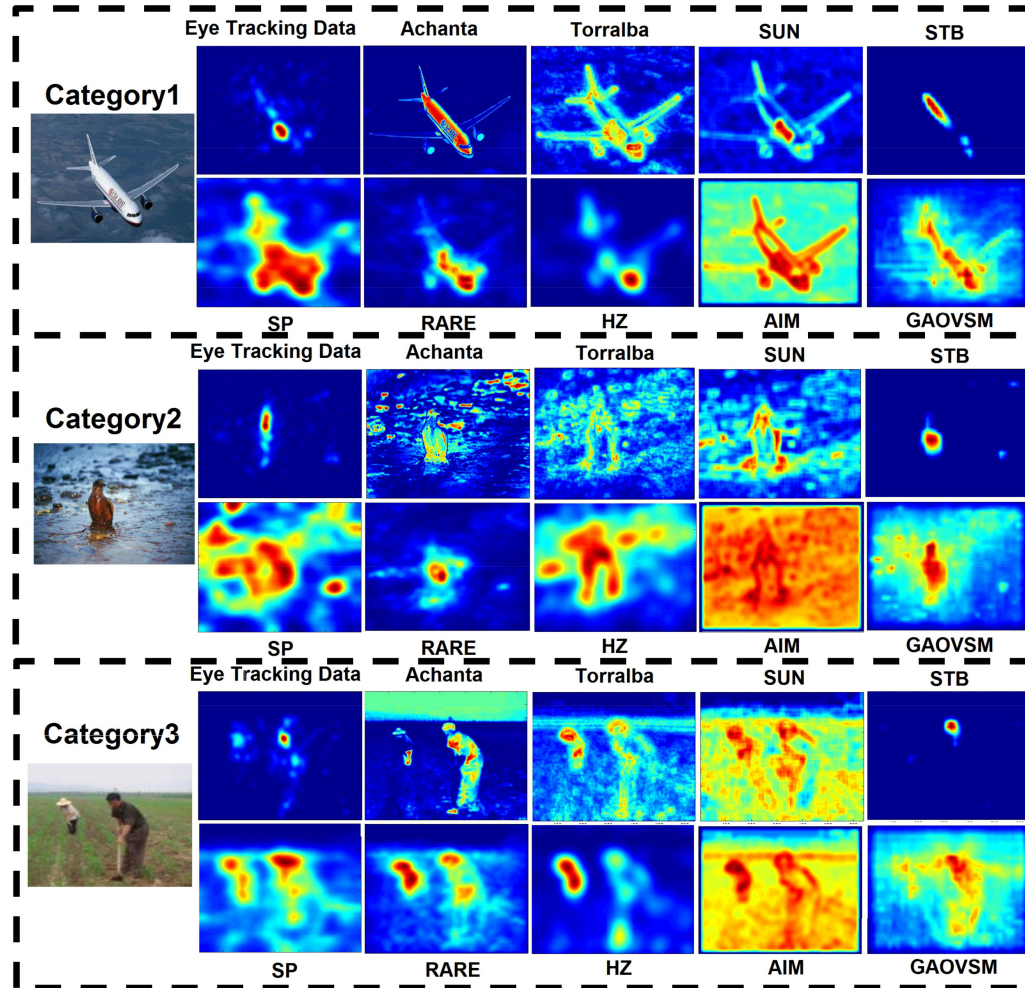


Figure 3.5: Qualitative comparison of visual attention methods for all categories.

to achieve qualitative improvements.

Notable improvements were shown by the proposed method as compared with the baseline method. The analysis of evolved solutions indicated that the prediction performance of the baseline orientation feature was greatly improved by the optimised parameters. The performance improvements are likely to be due to the proposed method weighting the orientation feature more heavily as compared with other features. This is in contrast to the baseline method, which employed unity weights for all features. The proposed GAOVSM performed better than several state-of-the-art methods and exhibited comparable performance to two high performance saliency methods, i.e. RARE and AIM.

This chapter was focused on the choice of system parameters in order to improve the performance of the traditional attention model on the task of human fixation prediction. The reason for choosing human fixation prediction as a task for evaluation is that the traditional model of attention is inherently better suited to fixation prediction, as it was originally designed for the task of human saccade generation.

The optimisation based attention method presented in this work is not expected to exhibit huge gains when tested on the task of salient object detection. This is anticipated due to the simple features employed by the traditional model of attention. Hence, the next chapter aims at developing new features that are more suited to salient object detection, and devising ways to incorporate them in a similar attention method.

## Chapter 4

# Spectral Matting Based Salient Object Detection

### 4.1 Introduction

During its infancy, visual saliency detection was limited to human fixation prediction. Recently the field has been extended to the identification of important regions in the scene containing a salient object [94], known as salient object or salient region detection. Salient object detection differs from the task of fixation prediction (discussed in chapter 3) in that the desired output is required a labelling of each pixel belonging to the salient object as foreground and the rest as background.

Studies of the human visual attention system have demonstrated that visual saliency is related to rarity, surprise and uniqueness of image regions. Such regions can be discriminated by their primitive signature features such as colour, texture and shape. Recently a plethora of techniques have been proposed that attempt to extract salient regions by computation of region-based contrast (in terms of primitive features) in a local or global fashion using either single or multiple scales [67, 91, 127, 147, 148, 157]. Such approaches initially segment the image into perceptually homogeneous elements before region based saliency computation. The saliency of

a region is then computed as a contrast with respect to any primitive feature, either by comparing a region with its surrounding superpixels [25], by computing the global contrast [101] or using a combination of both local and global contrast [157].

Before discussing the shortcomings of regions based approaches, a few important terminologies must be defined. The property of objects to be composed of multiple parts is termed as object composition. The term inappropriate annotation can refer to two undesired characteristics of a saliency map: 1) a saliency map falsely highlights background regions; 2) a saliency map highlights only some parts of the salient objects. The term non-uniform saliency assignment defines another undesired characteristic of a saliency result where notably different saliency values are assigned to different regions of the salient object.

Region based methods ignore important object properties such as object boundaries, shapes and composition resulting in inappropriate annotation and non-uniform saliency assignment. In an attempt to counteract these problems such methods often include multi-scale saliency maps or pixel level smoothing operations. However, after the first stage of processing using super-pixels based regions, it becomes cumbersome to retain the characteristics of proto-objects. These common problems associated with existing approaches are illustrated in Figure 4.1, along with the intuition that a technique is needed that can extract the complete salient object(s) by accurately separating it from the background.

Based on the above discussions, a technique is required that can assign uniform saliency inside object contours, has the ability to cover all distinct parts of the foreground salient object(s) and can effectively suppress background noise. It is noted that such a technique is based on the observation that foreground objects are characterised by their homogeneous nature, therefore uniform saliency inside foreground regions is desired. It is also acknowledged that for some applications, such as a medical tumour diagnosis, the heterogeneity of the foreground and the homogeneity of the

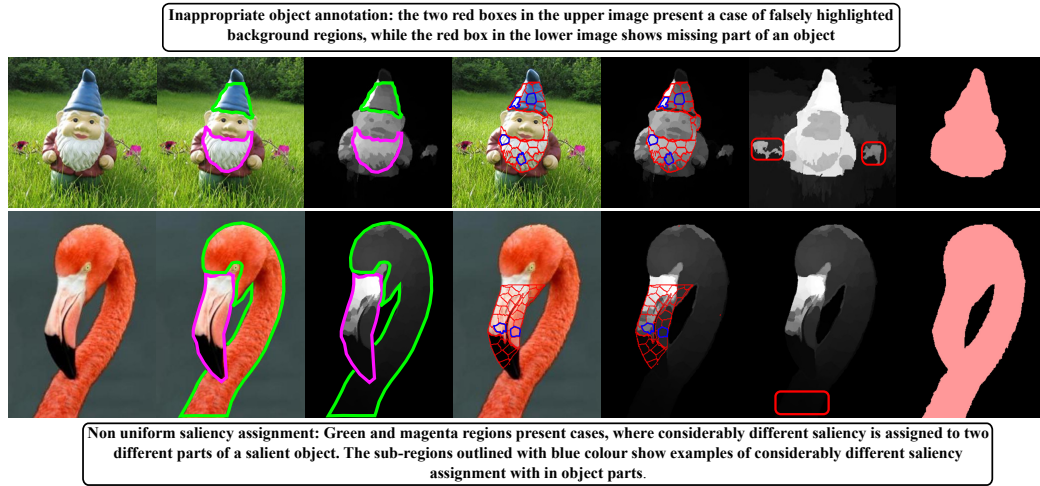


Figure 4.1: Illustration of inherent artifacts of the region based approaches. The first column presents the representative input images, columns 2-5 present examples of non-uniform saliency assignment, the sixth column depicts inappropriate object annotation and the last column shows the desired response. The saliency maps shown in this figure are computed by recent state-of-the-art region based approaches from the work of Zhu et al. [157] (column 5), that employs multivariate normal distribution termed as MND here, and the work of Yan et al. [147] (column 6) that proposed hierarchical saliency detection, termed as HS in this work.

background facilitates the classification of the tumour. Therefore, it will be an interesting future direction to investigate ways in which the desired uniform saliency based approach can be adapted to such tasks by changing the notion of foreground and background.

Digital image matting approaches seek effective foreground estimation to accurately segment the foreground from the background [87, 88]. These approaches model the observed image  $I$ , as a convex combination of foreground image  $F$  and a background image  $B$  by utilizing an alpha matte  $\alpha$ , such that  $I = \alpha F + (1 - \alpha)B$ . Therefore,  $\alpha$  defines an observed pixel as definite foreground when  $\alpha = 1$ , as definite background when  $\alpha = 0$  or mixed otherwise. For image matting,  $\alpha$  can assume values in the range  $[0,1]$ . However, if  $\alpha$  is constrained to only two values, i.e. 0 and 1, the matting problem condenses to the classical binary foreground/background segmentation problem, in which each pixel can either belong to the foreground or to the background. Even in its continuous form, the objective of matting is to reduce the number of mix pixels by investigating that whether they belong to the foreground or the background. Hence, the matting approaches aim to label all the pixels belonging to the foreground similarly as definite foreground, while ensuring similar labels for the background pixels, i.e. definite background. This is highly coupled with the task of salient object detection, in which the task is to label all the pixels belonging to the salient object(s) as salient and all the remaining pixels as background, whereas it is in high contrast with the region based salient object detection approaches that could assign different values to different pixels of the salient object (as an artifact of their processing). Based on the above discussion, it is arguable that whether the region based approaches more closely follow the actual formulation of the salient object detection problem as compared with the matting approaches.

One of the most promising approaches for digital matting employs matting components of a matting Laplacian matrix to accurately separate the foreground from the background [88]. The matting Laplacian is



a matrix that represents the pixels of an image using a graph in order to find important relations between them. The eigenvectors of the Laplacian matrix are employed to find partitions in the data with the objective of obtaining a single partition that separates the foreground from the background. A linear transformation of these eigenvectors yield the matting components, which contain the low level building blocks of semantically meaningful foreground objects. For each pixel of the image, every element of a matting component holds a value in the range  $[0,1]$  that represents the membership of that pixel (i.e foreground, background or mixed). However, the matting components span all the segments of an image including background segments. Hence, a method that can select the matting components containing only foreground information is of utmost importance for accurate segregation of foreground from background.

Despite the potential of benefiting a saliency application, the formulation employed by the image matting techniques have not been investigated in salient object detection literature. It is hypothesised that matting components have the ability to span all the distinct regions of the foreground salient object(s) and the capability of effectively suppressing the background noise. The hypothesis is based on prior works from the spectral matting literature [88, 87]. However, identifying those matting components, which are likely to contain foreground salient object(s) is a highly challenging task. Additionally, matting components of the matting Laplacian matrix are difficult to adapt for salient object detection, due to their high computational load.

#### 4.1.1 Chapter Goals

The goal of this chapter is to devise a technique that can quickly compute accurate matting components and employ them to construct effective foreground object saliency (through robust selection of foreground matting components) for salient object detection. The detailed objectives required

to achieve this goal are:

1. Devise a method for accurate computation of matting components with reduced computational time, thus making them adaptable to salient object detection.
2. To introduce novel matting component selection cues to identify the matting components that contain either a part or whole foreground salient object(s).
3. To introduce a method for coarse region of interest detection (i.e. a region with high probability of containing the salient object) to aid the matting component selection cues.

The anticipated outcome is a figure-ground segregation (FGS) method that can detect the whole salient object and suppress the unwanted background noise.

### 4.1.2 Chapter Organisation

The remainder of this chapter is organised as follows: Section 4.2 presents the implementation details of the proposed FGS method. Section 4.3 presents the design of experiments. Section 4.4 reports and discusses the results obtained. The final section presents a summary of important findings of this chapter.

## 4.2 Matting based Figure-ground Segregation (FGS) Method

The system model for the FGS method is presented in Figure 4.2. The final output of the FGS method is a foreground object saliency feature depicted by the symbol  $f^o$  (see section 4.2.4). The details of the individual components of the FGS method are discussed below.

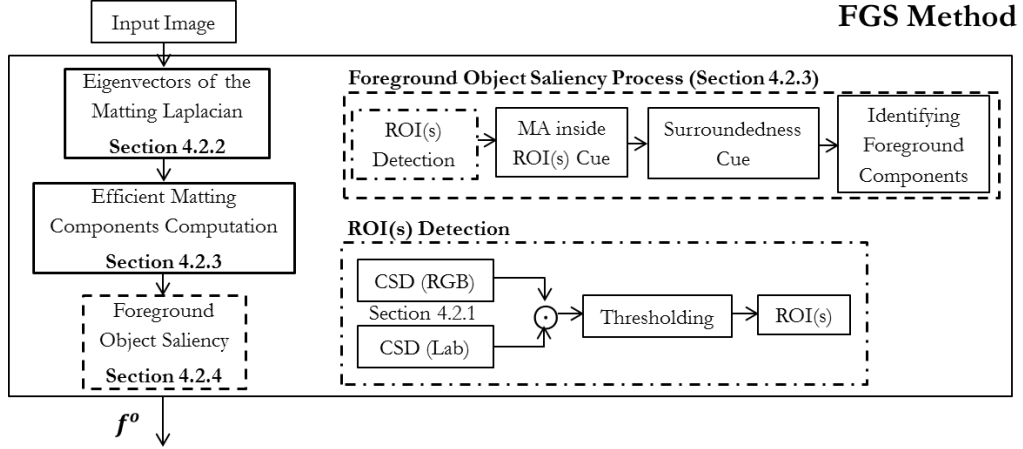


Figure 4.2: System model of the proposed FGS approach. The process depicted with a dot enclosed by a circle symbol in the ROI(s) detection block implies per element-wise multiplication of the CSD (RGB) and CSD (Lab) features.

### 4.2.1 Colour Spatial Distribution

The distribution of colours in space is a highly important cue in saliency detection. The idea is that a colour, which is widely distributed in an image is not likely to belong to a salient object [94]. The colour spatial distribution (CSD) feature is employed in the region(s) of interest ROI(s) detection step of foreground object saliency feature section 4.2.4.

The most suitable way of capturing the colour spatial distribution cue is to cluster colours of an image and model each colour as a component of a Gaussian Mixture Model (GMM) [26, 94, 157]. Thereby, a GMM based representation of the colours in the image is utilised in this work, where each pixel colour  $I_q$  is represented by several GMM components.

The GMM algorithm involves expectation maximization (EM) of the log-likelihood function. The stopping criteria for the EM algorithm is de-

defined as:

$$\left| \frac{\log(\mathbf{w}\mathcal{N}(\mathbf{I}|\boldsymbol{\theta}^t))}{\log(\mathbf{w}\mathcal{N}(\mathbf{I}|\boldsymbol{\theta}^{t-1}))} - 1 \right| < L_\tau, \quad (4.1)$$

where  $L_\tau$  is the log-likelihood threshold and  $\mathbf{w}$  is a vector containing prior probabilities of the GMM components. The parameter set  $\boldsymbol{\theta}^t$  includes the colour centres ( $\boldsymbol{\mu}^t$ ) and the covariance matrix ( $\boldsymbol{\Sigma}^t$ ) of the GMM components of the image  $\mathbf{I}$  for the current iteration  $t$ , while  $\boldsymbol{\theta}^{t-1}$  holds the centres and covariance matrix for the previous iteration. The log-likelihood threshold  $L_\tau$  is an important parameter as it influences the learned parameters of the GMM.

As directly employing GMM based clustering of colours ignores spatial correlation amongst individual components [26], a message-passing based clustering technique is employed based on such relationships [42], such as used in related works [26, 157]. The message-passing based clustering is employed to cluster the GMM components. This approach groups data points into clusters using message passing between data points and the preferred cluster centres. Real valued preferences are given as input to the affinity propagation approach, which specify preferred cluster centres. To associate a point to a cluster centre, similarities between data points must also be supplied as input to the message passing method. In this work, similarities for the data points (i.e. the GMM components) are obtained by finding pairwise correlation between the GMM components [26]. Preferences for cluster centres share a similar value in our implementation as all GMM components are supported to be potential cluster centres. The message passing process [42] clusters the GMM components based on their pairwise correlations resulting in more homogeneous clusters, without the need of defining the number of clusters beforehand.

The CSD feature  $f_{\text{csd}}$  for an image is computed as a measure of spatial variance of its colours (in the horizontal and vertical directions), weighted by the distance of its pixels from the image centre. The intuition behind this is that the colours that vary less in the spatial domain are less likely

to belong to the salient object. Similarly colours that are distant from the image centre are less likely to be salient.

### 4.2.2 Eigenvectors of the Matting Laplacian

The sparse matting components of the matting Laplacian are shown to be formed by linear combinations of the smallest eigenvectors of the matting Laplacian by Levin et al. [88]. Therefore, the smallest eigenvectors are computed in this work by designing the following affinity function to capture the structural information of an image [88]. Each element  $A$  for the affinity matrix  $\mathbf{A} \in \mathbb{R}^N$  is given as:

$$A = \frac{1}{|\mathcal{W}|} \sum_{p \in \mathcal{W}} \left( 1 + (\mathbf{I}_p - \boldsymbol{\mu}) \left( \boldsymbol{\Sigma} + \frac{\epsilon}{|\mathcal{W}|} \mathbf{U} \right)^{-1} (\mathbf{I}_p - \boldsymbol{\mu})^T \right), \quad (4.2)$$

where  $\mathcal{W}$  represents an image window centred around pixel  $p$ , consisting of a set of pixel coordinates. The cardinality of the set  $\mathcal{W}$  is  $|\mathcal{W}|$ ,  $\boldsymbol{\Sigma}$  is a  $3 \times 3$  covariance matrix in the window  $\mathcal{W}$  and  $\boldsymbol{\mu}$  is a  $3 \times 1$  vector containing the mean of colour values in the window  $\mathcal{W}$ . Given  $r$  as the radius of the window  $\mathcal{W}$ , the kernel size is given as  $|\mathcal{W}| = (2r + 1)^2$  and  $\mathbf{U}$  is a  $3 \times 3$  identity matrix.  $\mathbf{I}_p$  is a vector of image pixel values having the same size as  $\boldsymbol{\mu}$ . Given  $\mathbf{A}$ , the degree matrix is  $\mathbf{D} = \text{diag} \{d_{11}, \dots, d_{NN}\}$ , where each of its elements is given as  $d = \sum_{q,i} a_{q,i}^i$ ,  $a^i$  is the  $i$ th row of  $\mathbf{A}$  and  $q$  is an index to each element of the vector  $a^i$ . The matting Laplacian is then computed as  $\mathbf{L} = \mathbf{D} - \mathbf{A}$ . The smallest eigenvectors of the matting Laplacian are then computed.

### 4.2.3 Matting Components from Eigenvectors

The sparse matting components (introduced in the work of Levin et al. [88]) span all the fine-grained foreground and background regions of an image. The process of computing the sparse matting components from

the smallest eigenvectors of the matting Laplacian has a high computational load thus making its application difficult in saliency detection. The process of matting components computation is accelerated by neglecting the inappropriate ones. Additionally, image downsampling is employed to reduce the computational load. Moreover, identifying the foreground components from the set of all matting components is a highly challenging problem, as it requires finding an approximate location of the foreground object. The identification of the foreground components and their fusion results in a novel foreground object saliency (see section 4.2.4).

The matting components of  $\mathbf{L}$  can be recovered by a linear transformation of the eigenvectors of  $\mathbf{L}$ . Given the eigenvectors matrix  $\mathbf{V} = [\mathbf{v}^1, \dots, \mathbf{v}^O]$  of size  $N \times O$ , the goal is to find a set of  $O$  linear combination vectors to form  $\mathbf{G}$  that minimizes a non-convex cost described by Levin et al. [88] as  $\sum_k |\mathbf{m}^{(k)}|^\delta + |1 - \mathbf{m}^{(k)}|^\delta$ . Once  $\mathbf{G}$  is known the  $k$ th matting component can be computed as  $\mathbf{m}^{(k)} = \mathbf{V}\mathbf{g}^{(k)}$ , where  $\mathbf{g}^{(k)}$  is a row vector of  $\mathbf{G}$ . The above cost is referred to as the sum of sparsity scores and is a robust measure of the sparsity of the computed matting components.

The sum of sparsity scores cost is optimized using Newton's method as in [88] by computing a sequence of second-order approximations of the sum of sparsity cost given as  $\mathbf{V}^T |\mathbf{m}^{(k)}|^\delta + \mathbf{V}^T |1 - \mathbf{m}^{(k)}|^\delta$ . As introduced before, the matting components aim to assign a definite membership to each of their elements, which essentially means that they ideally desire to assign definite foreground or definite background to each element. Therefore, the objective of the optimisation process is to ensure that each element is close to a definite membership and the number of undefined (mix) elements is reduced. Higher order terms of the second order optimisations are approximated by the matting Laplacian and its eigenvectors. A reweighting procedure is applied to pull the entries of a matting component towards definite membership, i.e. 0 (or 1), in order to aid the optimisation process. This is achieved by employing the weighted eigenvectors and the eigenvalues of the matting Laplacian. Putting it all

together,  $G$  is updated as

$$G = \frac{V^T |m^{(k)}|^{\delta-2} + V^T |1 - m^{(k)}|^{\delta-2} + \ddot{p}}{2(V^T V_w + E)} \quad (4.3)$$

and the matting components matrix is updated as

$$M = VG. \quad (4.4)$$

In (4.3),  $\ddot{p}$  is the approximation of higher order terms given by  $\ddot{p} = \sum_{q,i} e^i_{q'}$ , where  $e^i$  is the  $i$ th row of  $P = V^T L$  and  $q$  is an index. The denominator in (4.3) constitutes the weighting term, where  $V_w$  is a weighted eigenvectors matrix. The weights are computed by linear combination of the  $k$ th matting component.

The optimal solution for the above optimisation problem can be obtained by binary matting components. However, as the matting components are restricted to be linear combination of eigenvectors, they can not hold all binary values in practice. As the goal of the optimisation process is to make the binary components as close to binary vectors as possible, the optimisation process amounts to iterative update of  $G$  and  $M$  using (4.3) and (4.4), with the number of iterations as the stopping criterion. The choice of the number for iterations is based on the sum of sparsity scores for the computed matting components. The optimisation process tends to reduce this sum in every iteration. The suitable number of iterations was searched empirically in this work as detailed below.

A total of 100 images were randomly sampled from the MSRA dataset [94] (see section 4.3). For each image, a total of 100 iterations were evaluated and the average of the sparsity scores (which is equivalent to normalized sum of sparsity scores) of all matting components were computed. The computed normalized sum of sparsity scores for all the images were averaged over the 100 iterations to obtain the mean sparsity score as depicted in Figure 4.3. The shaded area shows the standard deviation of the normalized sum of sparsity scores from the mean sparsity scores. Two important observations are noted from Figure 4.3: 1) the rate of change in

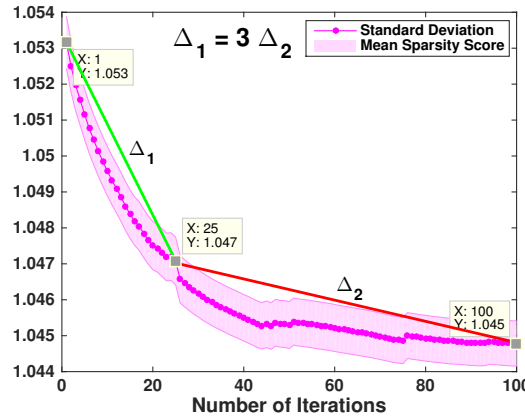


Figure 4.3: Convergence plot for the mean sparsity score of the matting components. The shaded area shows the standard deviation from the mean.  $\Delta_1$  and  $\Delta_2$  are the rate of change of mean sparsity score from iterations 1 to 25 and 26 to 100, respectively.

the mean sparsity score for the first 25 iterations (given by  $\Delta_1$ ) is triple the rate of change for the rest of the iterations (given by  $\Delta_2$ ). This encourages trading off the slow rate of change in the total sparsity score at the cost of saving additional computational time after 25 iterations; 2) additionally, substantial increase in the standard deviation from the mean sparsity score is observed after 25 iterations. This result suggests that the optimisation does not tend to remain effective for all images after 25 iterations. Therefore, the number for iterative updates is empirically set to 25 in this work.

To initialize  $M$  during the optimisation process, the  $k$ -means algorithm [41] is applied on the eigenvectors of the matting Laplacian. To speed up the initialization process, VL  $k$ -means algorithm from the VLFeat library<sup>1</sup> is employed. Default settings are used for VL  $k$ -means with  $k$ -means++ to initialize the cluster centres.

The optimisation process is sped up by neglecting those matting com-

<sup>1</sup><http://www.vlfeat.org/index.html>



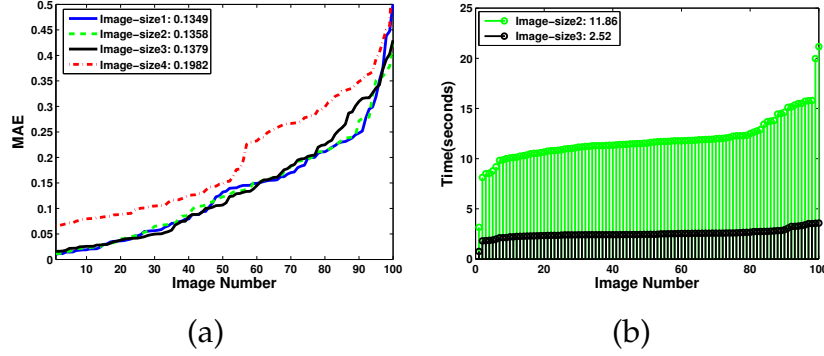


Figure 4.4: (a) Mean absolute error (MAE) for various image sizes. (b) Timing information (in seconds) for two selected image sizes. Image-size(1-4) represents original image downsampled by the ratios 1,0.7,0.3 and 0.1 respectively for each image-size. Both MAE and time are shown in sorted form. Mean MAE and time values are also shown. This figure is best viewed in colour.

ponents that have majority of their values in close vicinity of 0.5 as they negatively affect the optimisation. Such matting components make the sum of sparsity scores high, thereby slowing the optimisation process. They are detected by their characteristic high sparsity score, which is higher than 1.2, for the value of  $\delta$  used in our implementation.

In order to reduce the heavy computational load of eigenvectors and matting components, image downsampling is investigated. To show the affect of image downsampling on the computed matting components and the downsampling ratio used in this work, four representative image sizes are presented in Figure 4.4.

A total of 100 images were randomly sampled from the MSRA dataset to compute matting components for each image using various image sizes. To judge the quality of matting components produced corresponding to a particular size, the selected matting components were combined according to the procedure described in section 4.2.4 to form a foreground object saliency feature, whose quality reflects the quality of the matting compo-

nents. The quality of the foreground object saliency feature is measured by its mean absolute error (MAE) [113] given as

$$\text{MAE} = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H |S(x, y) - G(x, y)|, \quad (4.5)$$

where  $W$  and  $H$  are the respective width and height of the saliency  $S$  and ground truth  $G$ . MAE measures how well the saliency maps correctly identifies the background pixels and is a formal measure of figure-ground segregation quality. Therefore, it well suits the task at hand, which is to assess the figure-ground segregation ability of the matting components. The process of selecting and combining individual matting components (detailed in section 4.2.4) does not create any bias in our experiments as it is an independent operation common to all image sizes.

The plot on the left of Figure 4.4 shows the MAE for the representative image sizes (i.e. ratios: 1, 0.7, 0.3, 0.1), while the plot on the right shows the computational time in seconds for ratios 0.7 and 0.3 for the 100 representative images. Notably, the MAE for image ratios 1 and 0.3 are similar based on the average MAE, while 0.1 suffers pronounced increase of 43.7%. This advocates the use of ratios 0.7 and 0.3. However, based on the timing information, a noteworthy speed up of 4.7 times is achieved when using image ratio 0.3. Based on this observation, downsampling ratio of 0.3 is used in this work reducing the number of operations from  $N^2$  to  $0.3N^2$  for matting components computation.

#### 4.2.4 Foreground Object Saliency From Matting Components

The foreground object saliency is computed from the matting components. It aids the proposed FGS method to combat the undesired effects of the inappropriate object annotation and non-uniform saliency assignment that are common to region based visual saliency approaches. The proposed

foreground object saliency overcomes the problem of inappropriate annotation by employing the sparsity property of the matting components and taking assistance from the proposed foreground matting component selection cues. The problems of missing part of an object and non-uniform saliency assignment to object parts are tackled by ensuring maximum coverage of object regions. This is achieved by computation of accurate matting components that span all the foreground object regions.

To obtain the foreground object level saliency from matting components, an unsupervised technique is required that can sift through the matting components and identify the matting components that belong to the foreground object. The purpose of this sifting process is to ensure that all those matting components that contain accurately segmented foreground regions are selected and that all the noisy matting components are discarded. To achieve this, two newly designed cues are implemented, i.e. mean activity inside the region(s) of interest cue and surroundedness cue. The former cue ensures that only those matting components are selected that have mean activity inside the region(s) of interest greater than outside the region(s) of interest, while the latter ensures that no matting component is selected that contains a foreground region occupying more than one edge of an image. The foreground object saliency can then be computed as:

$$\mathbf{f}^o = \sum_k \beta^{(k)} \mathbf{m}^{(k)}, \quad (4.6)$$

where  $\beta^{(k)}$  (termed as foreground label) is the  $k$ th column of  $\beta$ , which is a  $K$ -dimensional binary vector. The  $i$ th location of the foreground label vector  $\beta$  is set to one for a foreground component and zero to specify a background component. Each element of  $\beta$  is computed by employing a mean activity inside regions of interest cue in conjunction with a surroundedness cue. The process of regions of interest detection is explained next.

### Region(s) of Interest Detection

Region of interest (ROI) detection has been employed in prior works [144, 93] to localize fine saliency regions. However due to outlier points of interest, background regions often get included in the region of interest, leading to a loss in performance. In contrast to previous work, only fine saliency regions are captured. First the RGB based colour spatial distribution feature (CSD)  $f_{\text{csd}}^{\text{R}}$  and Lab based CSD feature  $f_{\text{csd}}^{\text{L}}$  are computed using RGB and Lab based input images according to the process described in section 4.2.1. Next, the CSD feature maps are fused by per element wise multiplication ( $\odot$ ) to promote regions that are labelled as salient by both CSD features

$$f_{\text{csdC}} = f_{\text{csd}}^{\text{R}} \cdot f_{\text{csd}}^{\text{L}}. \quad (4.7)$$

The combined CSD feature  $f_{\text{csdC}}$  ensures that it only captures the regions that are voted to be salient by both the CSD maps. Afterwards,  $f_{\text{csdC}}$  is segmented to obtain the ROI(s).

The proposed  $f_{\text{csdC}}$  is generally robust to background noise and the foreground saliency feature is not highly sensitive to regions(s) of interest unlike previous approaches [144, 93]. Therefore, adaptive threshold based segmentation is employed in this work. A suitable value for the adaptive threshold that generalizes to unseen images is empirically sought, where the objective is to select the value that provides the best foreground object saliency performance. Based on past works [55, 3], the candidates for adaptive threshold values are set as multiples of the mean intensity of the  $f_{\text{csdC}}$  and the best value that corresponds to the best foreground object saliency performance is searched. The best value for the adaptive threshold was empirically found to be twice the mean intensity of  $f_{\text{csdC}}$  in this work.

### Mean Activity Inside Region of Interest Cue

The mean activity inside the ROI(s) ( $\mu_{\text{in}}$ ) and the mean activity outside the ROI(s) ( $\mu_{\text{out}}$ ) is computed as:

$$\mu_{\text{in}} = \frac{1}{|\mathbf{p}^{\text{I}}|} \sum_q \mathbf{p}_q^{\text{I}}, \quad \mu_{\text{out}} = \frac{1}{|\mathbf{p}^{\text{O}}|} \sum_q \mathbf{p}_q^{\text{O}}, \quad (4.8)$$

where  $\mathbf{p}^{\text{I}}$  represents the vector of pixel intensities inside the ROI(s) and  $\mathbf{p}^{\text{O}}$  is the vector containing pixels intensities outside the ROI(s).  $|\mathbf{p}^{\text{I}}|$  and  $|\mathbf{p}^{\text{O}}|$  are the number of elements in the vectors  $\mathbf{p}^{\text{I}}$  and  $\mathbf{p}^{\text{O}}$ , respectively. The mean activity inside the ROI(s)  $\mu_{\text{in}}$ , must be greater than the mean activity outside the ROI(s)  $\mu_{\text{out}}$  for a component to belong to the foreground.

### Surroundedness Cue

The Gestalt principle of surroundedness is employed for figure-ground segregation [109]. According to the principle of surroundedness: in ambiguous situations with two distinguished regions, when one region surrounds the other, the surrounded regions are always perceived as figures [109]. Zhang and Sclaroff [151] used the principle of surroundedness to produce an attention map from segmented colour channels of an image. They use the surroundedness principle in conjunction with a flood fill approach to condition the Boolean feature maps to obtain an attention map. Drawing on their approach [151], this work uses the Gestalt principle of surroundedness to identify matting components that belong to the foreground object.

To employ the surroundedness principle for foreground component detection, segmented matting components are used to capture their figure and ground proposals. According to the principle of surroundedness, the matting components that have a part of the figure present at their boundaries can not provide good figure-ground proposals. Specifically, it is noted that an object will rarely occupy more than one edge of an image. Based on this observation all the matting components having part of the

figure present on more than one of their sides are considered as outliers. The  $k$ th matting component is binarised as follows:

$$\mathbf{b}^{(k)} = \text{thresh}(\mathbf{m}^{(k)}, \theta), \quad (4.9)$$

where  $\theta$  is the segmentation threshold set as twice the mean intensity of the matting component. The threshold is set high to determine whether any part of the definite foreground is present at the boundary. Next, the boundary pixels membership sets  $\mathcal{B}^t, \mathcal{B}^b, \mathcal{B}^l, \mathcal{B}^r$  are computed, which contain all the non-zero pixels present at the components top, bottom, left and right boundaries respectively. Then, the boundary pixels membership sets are used for all the components to obtain a  $K$ -dimensional binary vector  $\boldsymbol{\eta}$ , where each of its elements are inspected to ensure that at least three of the boundary membership sets are empty for each component  $\mathbf{b}^{(k)}$  (i.e.  $\eta^{(k)} = 1$  iff at least three of the boundary membership sets are empty).

It is hypothesised that the surroundedness cue will not be highly beneficial in identification of foreground components if used in isolation, as it only helps in detecting outlier matting components (in cases where neither of the regions surrounds the other). It is anticipated that employing it in conjunction with the mean activity (MA) inside ROI(s) cue will greatly improve the overall salient object detection performance. Therefore it utilised as a complementary cue to the MA inside ROI cue, for determining if a component is to be selected as a foreground component or not.

### Identifying Foreground Components by Defining $\beta$

To detect the foreground components and form the foreground label  $\beta$ , the mean activity inside ROI(s) based selection is coupled with surroundedness based outlier detection. The foreground label for the  $k$ th matting component can be evaluated using MA inside ROI(s) and surroundedness cue as:

$$\beta^{(k)} = \begin{cases} 1 & \text{if } \mu_{\text{in}} > \mu_{\text{out}}, \eta^{(k)} = 1 \\ 0 & \text{otherwise.} \end{cases} \quad (4.10)$$

This process is repeated for all matting components to form  $\beta$ .

## 4.3 Design of Experiments

### 4.3.1 Dataset

The ImgSal database employed in Chapter 3 can not be used to evaluate salient object detection methods. Therefore, a database that can thoroughly evaluate the performance of salient object detection methods is required. **MSRA-B** [94] is a subset of the MSRA salient object dataset containing 5000 images collected mainly from the Internet. Recently Jiang et al. [68] provided pixel accurate segmentations for the whole database. MSRA includes images containing objects belonging to a wide variety of classes, while most of the images contain a single salient object. This dataset is widely used to measure the generalisability of methods due to the large number of available images. Due to its variety of image classes and extensive use for benchmarking the state-of-the-art, it is employed in this chapter for performance evaluations.

### 4.3.2 Parameter Settings

The dimension  $N$  of the eigenvectors matrix  $V$  and the matting components matrix  $M$  is defined as the number of pixels in the input image. Two important design parameters of the proposed FGS method are the number of smallest eigenvectors  $O$  and the corresponding number of matting components to be computed. During the initial experimentation, 10 matting components were found to provide reasonable coverage in terms of spanning the distinct regions of foreground object(s) for images drawn from the MSRA benchmark dataset. For this reason, 10 matting components were computed throughout this work. To obtain accurate matting components, the number of eigenvectors  $O$  must be greater than the number of matting components to be computed. After experimenting with

different values, the use of the 50 smallest eigenvectors were found to be a good choice for computing the matting components.

### 4.3.3 Evaluation Benchmarks

The standard benchmark for salient object detection, which performs fixed (naïve) thresholding of the saliency map at thresholds within the range  $[0, 255]$ , is employed in this work for evaluation of the performance [3]. Precision and recall are computed at each threshold to make a precision recall (PR) curve. The area under the PR curve (AUCPR) is then used as a measure of performance for comparing methods.

## 4.4 Results and Discussion

A quantitative and qualitative comparison of the proposed FGS approach with 10 state-of-the-art recently proposed saliency methods is presented. The methods include: MND [157], DSR [91], MR [148], MC [67], BSM [144], CS [45], LRK [123], SIA [26], HS [147] and DRFI [68]. These methods are included based on relevance to this thesis, recency and availability of their saliency maps.

Figure 4.5 shows the quantitative comparison of the proposed FGS approach with the state-of-the-art methods according to the fixed thresholding benchmark. It can be observed from the AUCPR scores that the proposed FGS method is ranked 3rd of the 10 state-of-the-art methods. It can also be observed that at lower thresholds and higher recall values, the proposed FGS approach maintains higher precision as compared to all the compared methods except DRFI. This demonstrates the ability of the proposed FGS approach to suppress background noise.

Figure 4.6, demonstrates the effectiveness of the proposed FGS approach as compared with the selected best performing state-of-the-art method. The best performing methods are selected based on their quantitative per-



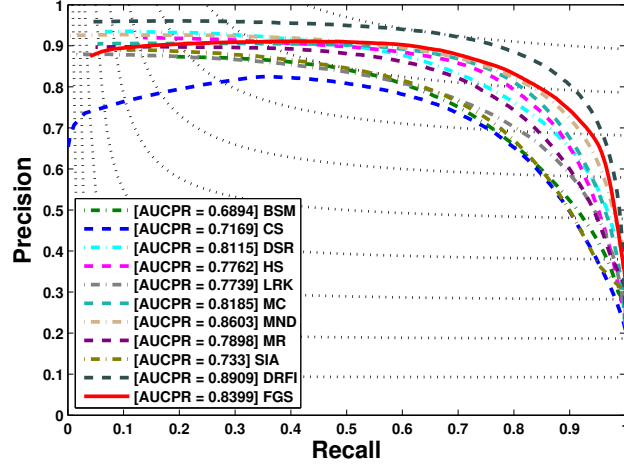


Figure 4.5: Quantitative comparison of the proposed FGS method with selected region based methods in terms of PR curves.

formance. The red annotations show the undesired artifact of inappropriate object annotation, which is a characteristic response by a few top performing state-of-the-art methods. DRFI includes unwanted background noise for the first two images, while the response of HS is corrupted with background noise for each representative image. The MND method effectively suppresses the background noise, however, it exhibits non-uniform saliency assignment for all representative images. In contrast to the top performing state-of-the-art methods, the proposed FGS method successfully suppresses the background and assigns uniform saliency inside object contours for all representative images.

#### 4.4.1 Discussion of FGS results

This section demonstrates that maximum coverage of object regions helped the proposed FGS approach in overcoming the problem of non-uniform saliency assignment. It also shows that effective selection of sparse matting components aided FGS to overcome the problem of inappropriate object annotation.

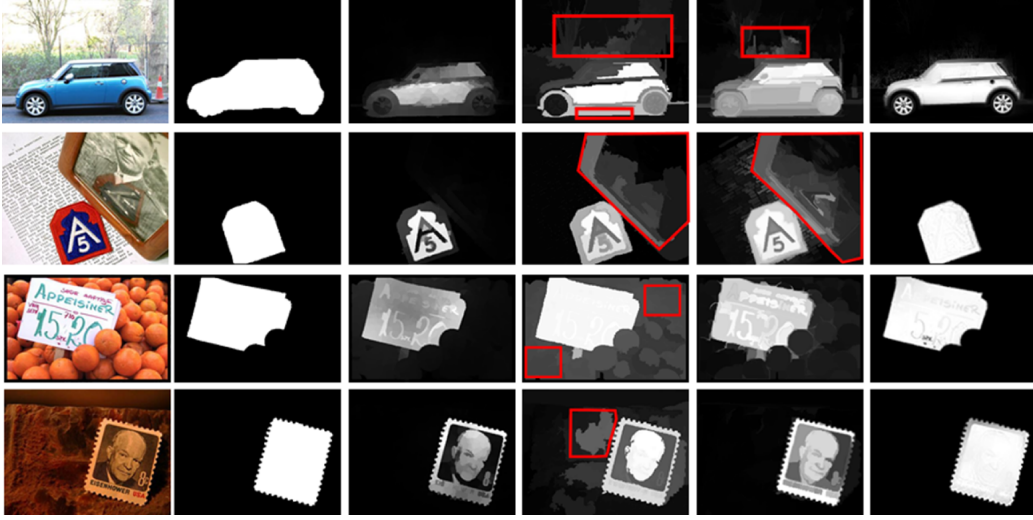


Figure 4.6: Visual comparison of the proposed FGS technique with selected state-of-the-art methods on representative images from the MSRA dataset. From left to right: image, ground truth, MND [157], HS [147], DRFI [68] and the proposed FGS method.

Figure 4.7 demonstrates the object coverage ability of the matting components computed according to the proposed approach. It can be observed that different parts of an object with different object properties are effectively covered by the object proposals obtained by the proposed matting components, resulting in uniform saliency assignment and appropriate highlighting of all foreground object parts.

Figure 4.8 demonstrates the procedure of neglecting unwanted noisy matting components and promoting sparser foreground components by the proposed matting component selection cues detailed in section 4.2.4. It can be observed that the selected components, shown inside green boxes, have greater mean activity inside the ROI(s) and also have surrounded foreground objects according to the definition of section 4.2.4. All the undesirable matting components with background noise are neglected by the selection cues. The matting component in the red box is neglected despite

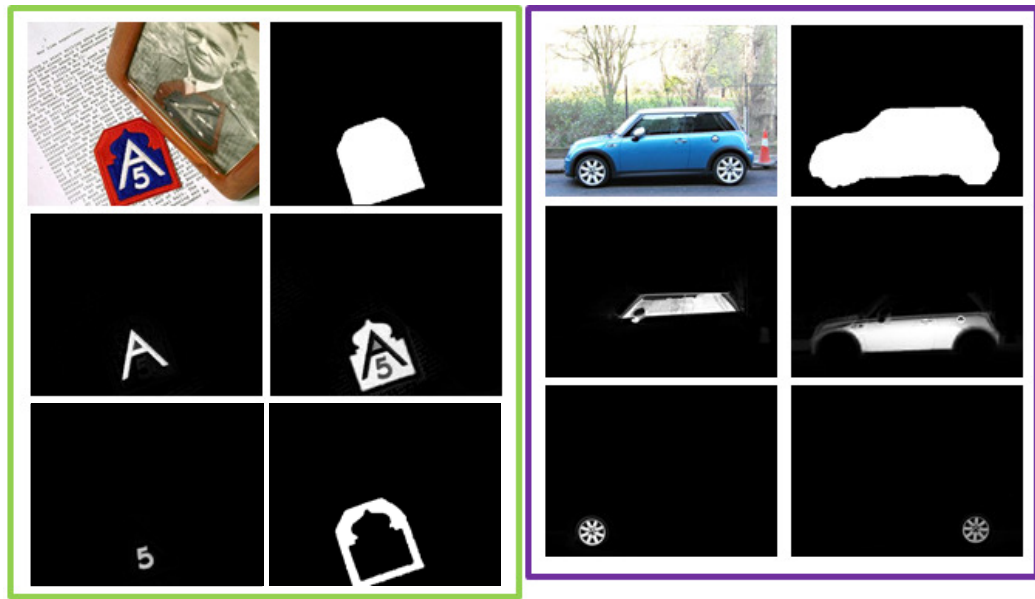


Figure 4.7: Object proposals by the matting components covering various parts of the foreground object. Only representative matting components are presented here. The colour boxes contain input image, desired ground truth and four representative matting components providing object proposals.

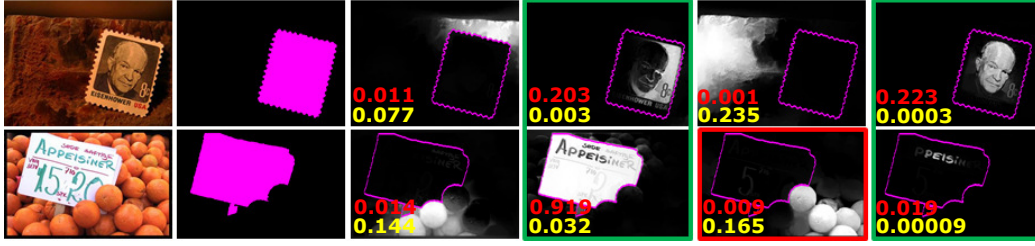


Figure 4.8: Selecting noise free sparse matting components by employing the proposed selection cues (section 4.2.4). The green boxes present the selected matting components, while the red box shows an unwanted matting component, which is deselected by both the cues. The figures in red depict the mean activity inside the ROI(s), while the figures in yellow show mean activity outside the ROI(s). The boundaries of the foreground objects are overlaid in pink for all matting components.

its sparse response, as it has greater activity outside the ROI and also does not satisfy the property of surroundedness.

## 4.5 Chapter Summary

The proposed FGS method was shown to be able to combat the undesired artifacts of saliency inherent to region based approaches, resulting in smoother and more pleasing saliency maps as compared with the region based state-of-the-art approaches. It was also shown that by suppression of the undesired background noise and through uniform saliency assignment to object parts, the proposed approach was able to improve upon several state-of-the-art approaches in terms of quantitative performance. Despite, this promising performance of the proposed FGS method, careful analysis of the FGS results revealed that in rare cases, noisy matting components are falsely selected by the proposed selection cues, thereby corrupting the FGS output. As the proposed selection cues are adaptive and unsupervised, they are limited in terms of robustness due to lack of



Figure 4.9: From left to right: input, ground truth (GT),  $f_{\text{csd}}^R$ ,  $f_{\text{csd}}^L$ , columns 5-7 show the first three smallest eigenvectors of  $L$  and column 8 shows  $f^o$ .

prior information or any supervisory input. Incorporating supervisory techniques or any prior knowledge can lead to high computational costs and is not desirable, given the heavy computational load of matting component computation.

On further analysis, it was revealed that in such scenarios, accurate saliency solutions can be obtained by the smallest eigenvectors of the matting Laplacian, as they are used to compute the matting components. Additionally, as the matting components exploit the colour information on a local scale using image windows, adding global colour features can aid the foreground object saliency in such scenarios. A representative example of such a case is presented in Figure 4.9, where one of the three smallest eigenvector of the matting Laplacian (columns 5-7) provides an accurate saliency solution and the global colour spatial distribution features in RGB and Lab space (columns 3-4) also provide good saliency solutions to aid the corrupted foreground object saliency. Further examples and analysis regarding the utility of eigenvectors and colour spatial distribution features in aiding the foreground object saliency feature is provided in chapter 5.

Based on the aforementioned hypothesis, the next chapter will evaluate the performance of the proposed FGS method when its final output response  $f^o$  is aided by the complementary smallest eigenvectors and global colour features.



## Chapter 5

# Improving the Figure Ground Segregation System By Learning Complementary Feature Combination

### 5.1 Introduction

The previous chapter introduced the FGS method, which was able to overcome the problems of inappropriate annotation and non-uniform saliency assignment by accurate computation and selection of matting components. Despite promising performance of the FGS method, a limitation in further performance improvement was found to be the noisy matting components that are falsely detected by the proposed unsupervised component selection cues. Potential solution to improve the component selection technique is to introduce supervised learning, but at the cost of additional computational time. As the computational overhead of the feature computation stage of the FGS methods is already high, it is not desirable to devise a technique that adds to the overall computation time.

Further analysis of the aforementioned problem revealed that the smallest eigenvectors of the matting Laplacians can contain accurate saliency solutions in such scenarios. As eigenvectors are required to compute the matting components, hence, they will not add any extra computational time to the method. Moreover, it was uncovered that adding global colour information to the local colour information exploited by the matting components can benefit the foreground object saliency feature. Favourably, as these features are already computed as part of the FGS method, they add no overhead to the computational time of the method.

The inclusion of multiple features in the proposed approach reflects the need for considering feature importance during combination (see section 5.2.1). Moreover, experimentation with a range of parameters (for details, see section 5.2.1) revealed that the parameters related to the feature computation process considerably affect the salient object prediction capability of the features. To improve upon the salient object detection performance of the FGS method, important parameters must be searched in conjunction with feature importance weights. Hence, the problem at hand can be formulated as an optimisation or search problem, where the parameters of the feature computation process and feature weights must be searched with the objective of improving the salient object detection performance.

Most methods in the literature employ *ad hoc* feature combination schemes for salient object detection [157, 101, 52]. These integration rules are fragile with poor generalizability due to their *ad hoc* tuning. A few learning-based salient object detection approaches have been proposed in the literature [133, 13, 72], which learn feature importance by formulating the salient object detection problem as a classification problem. These techniques are generally classification based methods, therefore parameter search can not be easily incorporated into their formulation. The recent work by Jiang et al. [68], on discriminative regional feature integration (DRFI), employs regression to automatically select and integrate features. However, in sce-



narios where feature performance is heavily dependent upon important feature related parameters, it is not feasible to incorporate joint optimisation of parameters into the automatic feature integration process of DRFI.

As the decision variables have bound constraints and a few decision variables are integers, the optimisation problem condenses to optimisation of a performance based objective function along with constraint handling for decision variables. The majority of traditional constrained optimisation methods [12] transform the optimisation problem into a convex equivalent [16] and also require domain knowledge. As the integer and bound constraints make the optimisation problem non-convex, the latter class of methods may not be directly adapted for this task. Hence, a system is needed that can perform efficient optimisation in a large search space, does not require prior knowledge of the search space and has the ability to encode performance metrics as its objective function. Amongst other optimisation approaches, a Genetic Algorithm (GA) is a competitive search technique in large search spaces. Furthermore, it does not require any prior knowledge of the search space and has the ability to handle integer and bound constraints on decision variables. Moreover, the GA has the ability to encode salient object performance metrics directly as its objective function.

### 5.1.1 Chapter Goals

The goal of this chapter is to devise a method for fusion of the foreground object saliency ( $f^o$ ) (see chapter 4) with the smallest eigenvectors of the associated matting Laplacian and the CSD features, by joint optimisation of the feature parameters and feature importance. The following objectives are set to meet this goal:

1. To investigate whether the proposed FGSopt method can benefit the foreground object saliency  $f^o$ , by improving its salient object detection performance.

2. To develop a baseline unoptimised method that uses human encoded parameters and assigns equal importance to all features. It will then be compared with the FGSopt method to investigate whether it is the optimised parameters that benefit  $f^o$  or solely the unoptimised complementary features.
3. Quantitative and qualitative comparison of the proposed FGSopt approach with the state-of-the-art learning and non-learning based salient object detection methods.

### 5.1.2 Chapter Organisation

The remainder of this chapter is organised as follows: Section 5.2 provides the implementation details for the proposed FGSopt method. Section 5.3 presents the design of experiments. Section 5.4 presents and discusses the results obtained. The final section presents an overview of important findings of this chapter.

## 5.2 Method

The system model for the proposed FGSopt method is depicted in Figure 5.1. The FGSopt approach has two operating modes, namely offline optimisation and saliency computation. During the offline optimisation mode, the important parameters of the feature computation and weighting process are searched by a Genetic Algorithm and stored in an optimal parameter vector  $\lambda$ . In the saliency computation mode,  $f^o$  is fused with additional complementary features, i.e. the smallest eigenvectors of the matting Laplacian and colour spatial distribution (CSD) according to the learned parameters  $\lambda$  to compute the final saliency  $s_f$ . The implementation details of  $f^o$ , smallest eigenvectors of the matting Laplacian and the CSD features are discussed in the previous chapter.

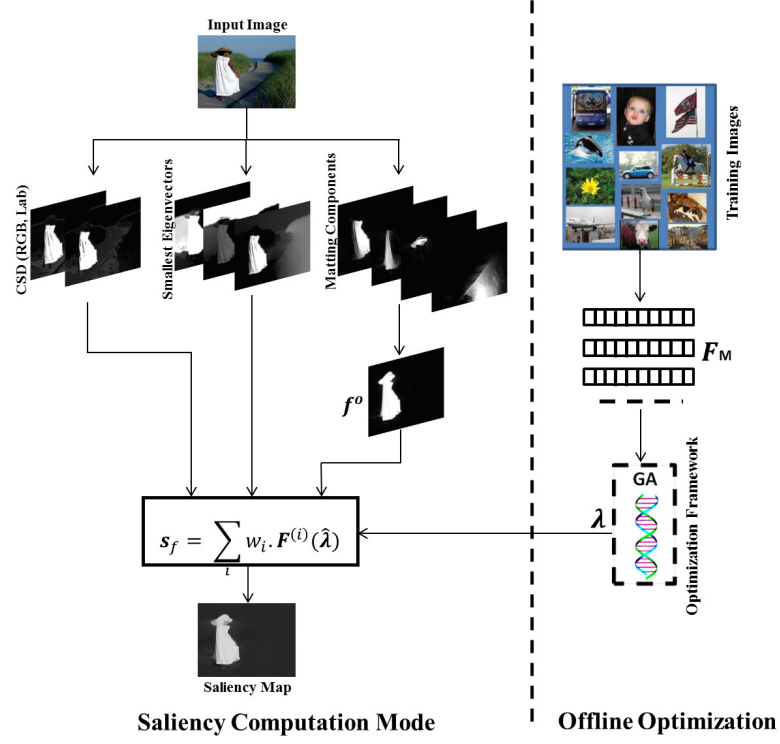


Figure 5.1: System model of the proposed FGSopt approach. Complementary features are computed using training images to form a feature matrix  $F_M$ , which is fed into the optimisation block. After offline optimisation, the learned parameter set  $\lambda$  is used to compute the final saliency map for all images in the saliency computation mode.

### 5.2.1 Optimisation Framework for Feature Combination

The visual results of the proposed saliency features using representative images are shown in Figure 5.2 to illuminate the rationale behind learning the feature importance during feature fusion. From the top row image it can be observed that the second smallest eigenvector of  $L$  and  $f^o$  better capture the salient object (i.e. the jumping boy) as compared with the other features and hence should be given more importance during the feature combination process. Conversely, for the second row image, the Lab based CSD feature  $f_{\text{csd}}^L$  and the second smallest eigenvector of  $L$  produce comparatively better saliency maps and therefore, should be weighted highly as compared with the other features. This necessitates the task of weighting features to quantify their relative importance. A wide class of methods have addressed this issue by learning the optimal weights [133, 158, 13, 72]. *While these methods achieve statistically good results on publically available benchmarks, they only focus on optimising feature combination neglecting the optimisation of bottom-up parameters.*

Levin et al. [88] used human encoded values for the parameters involved in bottom-up computation such as window size  $r$  and the regularization term  $\epsilon$  in (4.2). However the quality of segmentation is a function of window size  $|\mathcal{W}|$  and varying  $|\mathcal{W}|$  will affect the captured similarity and proximity of the neighbouring pixels in the window [88]. Also,  $\epsilon$  controls the trade-off between noise and smoothness in the final solution and affects the quality of the resulting eigenvectors in terms of their ability to capture true image segments [88]. To observe the effect of varying  $r$  and  $\epsilon$  on the quality of segmentation (at the first stage of foreground object saliency computation), 10 images from the MSRA-B dataset [94] were sampled and their corresponding mean absolute error (MAE) [113] profiles plotted in Figure 5.3. The MAE profile for varying  $r$  with a fixed  $\epsilon$  is shown on the left (top row) and the profile for varying  $\epsilon$  with fixed  $r$  is shown on the right (top row) of Figure 5.3. To generate these profiles for each image,  $n_v$  smallest eigenvectors of  $L$  were used. The average MAE



Figure 5.2: Rationale for learning feature importance during feature combination. From left to right: input, ground truth (GT),  $f_{\text{csd}}^{\text{R}}$ ,  $f_{\text{csd}}^{\text{L}}$ , columns 5-7 show the first three smallest eigenvectors of  $L$  and column 8 shows  $f^o$ . It can be clearly observed that different features obtain different level of accuracy in capturing the salient object. Hence, the technique must assign different importance to features having high accuracy for a particular image as compared to the features that obtain lower accuracy. It is also noted that individual feature accuracy also varies on image-by-image basis, therefore the relative importance of features must be learned for effective feature combination on unseen images.

and the standard deviation from mean MAE (depicted by the error bars) for specific values of  $r$  and  $\epsilon$  are shown. It can be observed that varying  $r$  and  $\epsilon$  in isolation (while one of them fixed) causes considerable variance in the quality of eigenvectors for each image and significant variance in MAE can be observed when considering all the images. A general trend for the average minimum MAE can be seen at  $r = 2$  and  $r = 3$  for the window radius profile. Similarly the lowest average minimum MAE is recorded for  $\epsilon = 1e - 4$  and  $\epsilon = 1e - 5$  for the regularization parameter profile.

However, it is not plausible to generalise these values over the whole set of images. Optimal values for these parameters that generalise over the whole dataset are therefore needed. To show the effect of simultaneously varying  $r$  and  $\epsilon$  over a grid of values, the MAE profile (bottom row of Figure 5.3) for the top row image (jumping boy) in Figure 5.2 is also plotted. As can be seen, this is not a simple unimodal trade-off surface. Although locating the minimum in this case is not cumbersome, the generalization

over multiple images would not be practical. Hence, the suitable values for these parameters of the bottom-up process must be searched in conjunction with the feature importance to improve performance.

### Formulation of the Optimisation Problem

Given a set of training examples  $\mathcal{T} = \{\mathbf{t}^{(1)}, \mathbf{t}^{(2)}, \dots, \mathbf{t}^{(Q)}\}$  with the corresponding annotations  $\mathcal{A} = \{\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \dots, \mathbf{a}^{(Q)}\}$ , the goal is to learn a parameter vector  $\boldsymbol{\lambda} = [r, \epsilon, n_v, L_\tau, \mathbf{w}]$  by minimizing the difference between the predicted saliency set  $\mathcal{S} = \{\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \dots, \mathbf{s}^{(Q)}\}$  and the annotations  $\mathcal{A}$  in terms of the positive predictive value (precision)  $\hat{p}$  and the hit rate (recall)  $\hat{r}$ . Given precision  $\hat{p} = \{p_1, p_2, \dots, p_n\}$  and recall  $\hat{r} = \{r_1, r_2, \dots, r_n\}$  on a set of thresholds  $\mathbf{t}_a$  such that  $r_1 \leq r_2 \leq \dots \leq r_n$  holds, let  $p_i^{\max}, p_i^{\min}$  be the largest maximum and minimum sample precision values corresponding to  $r_i$ , respectively. The objective function maximises the area under the PR curve (AUCPR) by minimizing the following:

$$\boldsymbol{\lambda} = \arg \min_{\boldsymbol{\lambda}} \left( 1 - \sum_{i=1}^{n-1} \frac{p_i^{\min} + p_{i+1}^{\max}}{2} (r_{i+1} - r_i) \right). \quad (5.1)$$

The thresholds are defined as  $\mathbf{t}_a = [0 \dots 255]$  according to the benchmark reported in [3]. For computational efficiency,  $\mathbf{t}_a$  is downsampled using a step size of four. During downsampling it is ensured that the samples are enough for a reasonable approximation of area under the curve using trapezoidal estimation [1]. The accuracy of area under the curve approximation may have pronounced effects on instances that are very close in the PR operating space. Given a thresholded saliency map  $\hat{\mathbf{s}}$  and the corresponding annotation  $\mathbf{a}$ , the precision for the  $i$ th threshold is computed according to (5.2).

$$\hat{p}(t_i) = \frac{\sum_q \mathbf{a}_q \cdot \hat{\mathbf{s}}_q(i)}{\sum_q \mathbf{a}_q \cdot \hat{\mathbf{s}}_q(i) + \sum_q (1 - \mathbf{a}_q) \cdot \hat{\mathbf{s}}_q(i)} \quad (5.2)$$

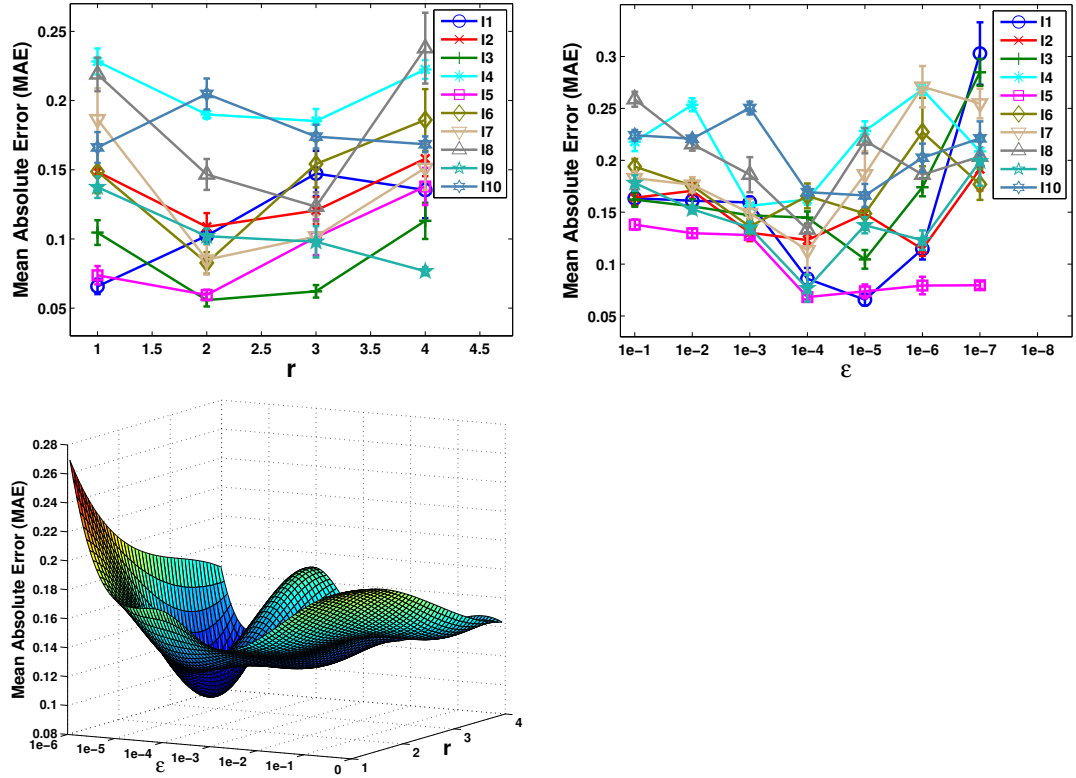


Figure 5.3: Rationale for learning important parameters of bottom-up saliency computation. Top row: effect of  $\mathcal{W}$  and  $\epsilon$  on the eigenvectors of  $L$ , bottom row: combined effect of jointly varying  $\mathcal{W}$  and  $\epsilon$  for the representative jumping boy image in top row of Figure 5.2. I1-10 represents the image number here for the selected representative images. This figure is better viewed in colour.

The recall for the  $i$ th threshold is given as:

$$\hat{r}(t_i) = \frac{\sum_q \mathbf{a}_q \cdot \hat{\mathbf{s}}_q(i)}{\sum_q \mathbf{a}_q}, \quad (5.3)$$

where  $\hat{\mathbf{s}}_q(i)$  represents a segmented saliency map at the  $i$ th threshold in (5.2) and (5.3).

### Genetic Algorithm (GA) Implementation

A real-coded Genetic Algorithm (R-GA) is employed to model the optimisation problem given by (5.1). The proposed GA algorithm creates a random initial population that is drawn from a uniform distribution, such that it satisfies the constraints on integer variables. A real-coded representation is used to encode the population and rounding is used to ensure integer constraints. The bounds on the decision variables are based on prior knowledge of the sensible variable ranges. The tolerance on the average change in fitness value along with the number of generations define the termination criteria for the GA.

$r$ ,  $\epsilon$  and  $n_v$ , are encoded as integer variables in the proposed GA based optimisation framework.

$$\begin{aligned} r &\in \mathbb{Z} : r^L \leq r \leq r^U \\ \epsilon &\in \mathbb{Z} : \epsilon^L \leq \epsilon \leq \epsilon^U \\ n_v &\in \mathbb{Z} : n_v^L \leq n_v \leq n_v^U. \end{aligned} \quad (5.4)$$

$L_\tau$  from (4.1) is encoded as

$$L_\tau \in [1 \times 10^{-3}, 9 \times 10^{-3}] \subset \mathbb{R}, \quad (5.5)$$

where the bounds are based on prior information from past work [94].

Finally the weights of the features  $\mathbf{w} \in \mathbb{R}$  are encoded in the following range:

$$\mathbf{w} \in [0, 1]. \quad (5.6)$$



For each saliency output  $s$ , its segmented version  $\hat{s}$  is compared with the corresponding ground truth annotation  $a$  to compute the respective  $\hat{p}$  and  $\hat{r}$ . The fitness for each  $s$  is then computed using the objective function in (5.1).

Binary tournament selection [47] is used to choose a pool of quality solutions by exerting high selection pressure. In combination, parent-centric recombination operators (i.e. Laplace cross-over operator and power mutation [32]) are employed in the anticipation that placing children in proximity with the parents can provide us with good solutions throughout the evolutionary process.

For the selection, crossover and mutation operators, the implementations described in chapter 4 are again used.

### 5.2.2 Final Saliency Computation

To compute the final saliency output  $s_f$  for an image, all the computed features are concatenated into a feature matrix  $F_M = [f_{\text{csd}}^R, f_{\text{csd}}^L, V_{n_v}, f^o]$ , where  $V_{n_v}$  contains  $n_v$  smallest eigenvectors of  $L$  as its columns. Then the final saliency  $s_f$  is computed as

$$s_f = \sum_i w^{(i)} \cdot f^{(i)}(\hat{\lambda}), \quad (5.7)$$

where  $w^{(i)}$  is the learned weight for the  $i$ th feature and  $f^{(i)}$  represents the  $i$ th column of  $F_M$ .

## 5.3 Design of Experiments

### 5.3.1 Datasets

As this chapter introduced the first method for joint learning of feature parameters and importance for salient object detection, several benchmark datasets were employed for thorough evaluation of the method. This was

to ensure that the performance is truly reflective of wide variations that can occur in the data.

The **MSRA-B** [94] dataset is employed in this work due to the reasons explained in chapter 4. That chapter also discussed specific details of the dataset.

**SED1** [9] is a subset of the segmentation evaluation database containing 100 single salient object images. Pixel accurate segmentations from up to three different users are provided in this dataset.

**SED2** [9] is also a part of the segmentation evaluation database having a total of 100 two salient object images with pixel accurate annotations similar to SED1. Due to two salient objects per image, it is difficult for methods to achieve optimum results on this dataset.

**SOD** [103] is a subset of the Berkeley Segmentation Database [102] containing 300 images. Scenes containing multiple salient objects with complex object and background appearance makes it a difficult dataset for most saliency methods. Seven subjects were asked to label the foreground object, where each subject can label more than one salient object in an image. A confidence score is provided for each labelled object, however a single foreground mask is not provided by the authors as the ground truth [103]. Thereby, the approach used in [136] is followed to generate the pixel accurate ground truth masks.

The **PASCAL-VOC2012** dataset (named as VOC) is a collection of images used in the PASCAL-VOC challenge since 2007. It has a total of 11,540 images. This work employs 2,913 images out of the total images, which have segmentation ground truth available. It contains the most difficult images for salient object detection having objects of more than 20 image classes with multiple salient objects. The pixel accurate ground truth segmentations provided for the object detection task are used in our evaluations.

A total of 2500 images are randomly sampled from the MSRA and the VOC datasets as training examples for the proposed optimisation frame-

work. Test performance is evaluated using all remaining images.

### 5.3.2 Parameter Settings

After searching for suitable values, the population size for the GA was set to 200 and the number of generations to 50. A crossover fraction of 0.8 and mutation rate of 0.01 were found to be suitable for the experiments with four elite individuals retained each generation. Optimal values for the parameters of the bottom-up process ( $r, \epsilon, n_v, L_\tau$ ) are searched during the optimisation procedure. The optimal values found after optimisation are  $r = 3, \epsilon = 1e - 6, n_v = 6$  and  $L_\tau = 7 \times 10^{-3}$ , which are employed to compute saliency for all unseen images.

### 5.3.3 Evaluation Benchmarks

Two standard benchmark methods as reported in [3] for saliency segmentation are employed during the performance evaluation. The first benchmark is the same as employed in chapter 4. It performs fixed (naïve) thresholding of the saliency map to generate a precision recall curve.

The second benchmark employs adaptive thresholding of saliency maps. According to [3], the image is segmented by using the mean shift segmentation algorithm [30] and only those segments are retained in the binary segmented saliency map for which the saliency value is higher than twice the mean saliency of the saliency map. Average precision and recall are computed by comparing the segmented binary map with the corresponding ground truth. Average F-measure is then computed according to Achanta et al.[3] as:

$$F_\beta = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \quad (5.8)$$

Similar to [3],  $\beta = 0.3$  is employed to give more importance to precision as compared with recall.

## 5.4 Results and Discussion

This section primarily compares: the performance of the proposed FGSopt method with the FGS method, then with the baseline unoptimised method to gauge any performance improvements and finally presents a rigorous comparison of the proposed FGSopt method with 10 state-of-the-art methods including MND [157], DSR [91], MR [148], MC [67], BSM [144], CS [45], LRK [123], SIA [26], HS [147] and DRFI [68]. These methods are included based on relevance to the proposed work, recency and availability of their saliency maps.

### 5.4.1 Comparison with the FGS method

Figure 5.4 shows the feature level performance as compared with the combined performance of the proposed FGSopt method. It can be noted that the proposed FGSopt method considers the performance gaps between individual features to improve upon the FGS performance. By optimally combining the FGS output with the complementary features, the overall performance of the FGS method is improved. The FGSopt method also improves upon FGS performance in the F-measure space as can be observed by the F-measure contours on the PR curve in Figure 5.4.

### 5.4.2 Comparison with the Baseline Unoptimised method

This section investigates the effects of the optimised method parameters on the overall saliency detection performance of the proposed FGSopt method. For comparison, a baseline method is implemented that employs human encoded values for all the variables in the parameter vector  $\hat{\lambda} = [r, \epsilon, n_v, L_\tau]$ . The weights vector for the baseline method is populated with ones for all features to assign equal importance to all features. The feature and method level comparison of the proposed FGSopt method with the baseline method is presented in Figure 5.5 and Table 5.1. Fig-

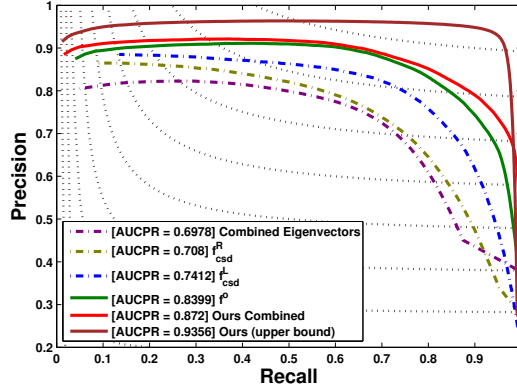


Figure 5.4: Comparison of the feature level performance with the overall proposed FGSopt method.

Figure 5.5 (a) presents the method level comparison of the proposed method with the unoptimised baseline method, while Figure 5.5 (b), (c) and Table 5.1 show the feature level performance comparison of both the methods.

For the smallest eigenvector features, the optimised parameters were found to suppress the background noise as compared to the baseline features that employ human encoded values. Representative examples of this behaviour are shown in Figure 5.5 (b) for the second smallest eigenvector of the matting Laplacian. The reason behind this behaviour is that the baseline eigenvector features tend to divide the image into many distinct clusters, whereas the optimised eigenvector features tend to minimize the number of clusters to only two clusters, i.e. the foreground and the background. This is clearly suggested by the top row flowers image in Figure 5.5 (b). Moreover, the quantitative performance evaluation shown in Table 5.1 on the MSRA dataset confirm the superior performance of the proposed optimised eigenvector features over the baseline unoptimised features.

Both the optimised RGB and Lab based colour spatial distribution features show minimal improvements over their baseline counterparts. This

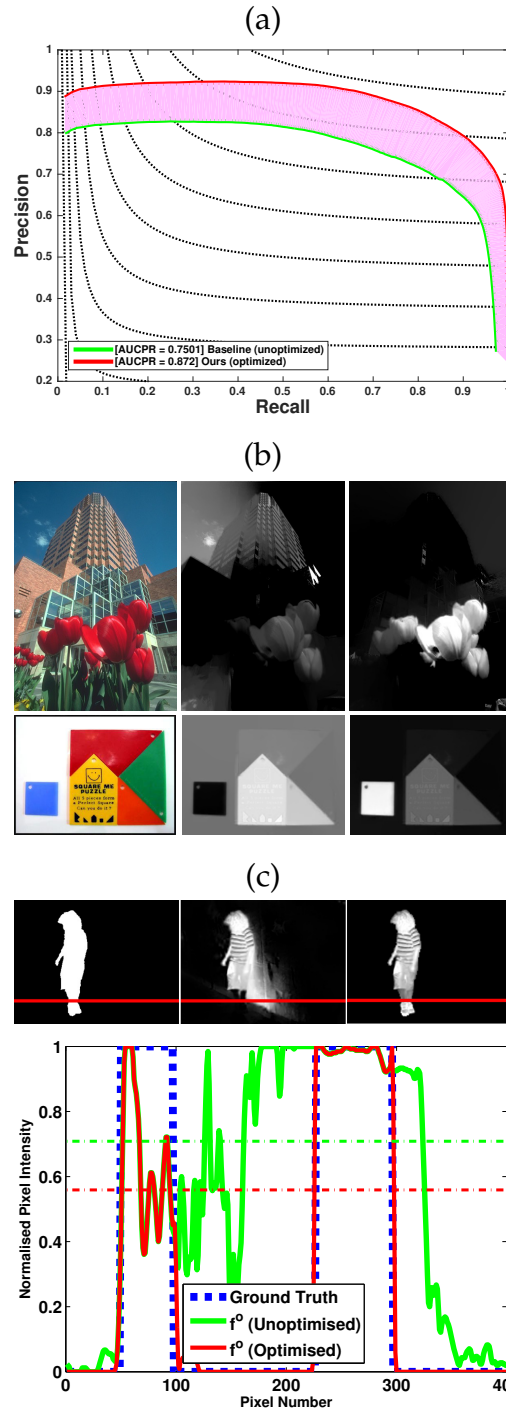


Figure 5.5: Comparison of the proposed FGSopt method with the baseline unoptimised method. (a) PR curves for the baseline and the proposed FGSopt method on the MSRA dataset. The shaded region signifies the performance improvement. (b) Effect of parameter optimisation on the second smallest eigenvector of the matting Laplacian. (c) Effect of parameter optimisation on  $f^o$ .

AUCPR	Baseline	Proposed	Percentage increase
<b>Combined eigenvectors</b>	0.6719	0.6978	3.7%
<b>CSD (RGB)</b>	0.708	0.715	1%
<b>CSD (Lab)</b>	0.7412	0.7510	1.3%
<b>Foreground object saliency feature</b>	0.8042	0.8399	4.4%

Table 5.1: Feature level performance enhancements by the optimised parameters on the MSRA dataset.

result suggests that the log-likelihood threshold does not greatly affect the saliency detection performance of the CSD feature when used in conjunction with complementary features.

The first row Figure 5.5 (c) shows the ground truth, baseline foreground object saliency and the optimised foreground saliency feature of a representative image. The second row of Figure 5.5 (c), shows the one dimensional (1D) profile of the images presented in the first row, where the red line overlayed on top row images shows the slice used to generate the 1D profiles. The red and green horizontal dotted lines plotted on the 1D profile depict the respective thresholds of the proposed and unoptimised feature maps, computed as twice their mean intensity as per the evaluation benchmark [3]. It can be clearly observed from the top row results that the proposed  $f^o$  feature suppresses the unwanted background noise present in its baseline unoptimised counterpart. This result is also reiterated by the 1D profile, where the proposed  $f^o$  more closely matches the ground truth profile as compared to its baseline unoptimised counterpart. The superior performance of the optimised  $f^o$  as compared to the baseline feature is also confirmed by considerable performance enhancements of 4.4%.

Figure 5.5 (a) shows the substantial improvement of 16.2% achieved

by the optimised parameters of the proposed FGSopt method in comparison to the baseline method. Considering the 99% confidence limits of the multiple runs of the FGSopt method, the performance improvements range from 12.4% to 20%. This is a reasonable result considering the individual feature level enhancements obtained due to the optimised features and also taking into account the important factor that all the features are given equal importance in the baseline method. Due to equal importance assignment to all features in the baseline method, highly conspicuous feature maps are suppressed by noisy features, thereby affecting overall generalization.

### 5.4.3 Comparison with the State-of-the-art methods

#### Segmentation Using Fixed Thresholding

The saliency maps are segmented using the fixed thresholding benchmark (section 5.3.3) to compute the PR curves presented in Figure 5.6. Figure 5.6 (a)-(e) show the PR curves of methods on individual datasets, while Figure 5.6 (f) shows the average performance of methods on all the datasets. In Figure 5.6 (f), the average performance of MND is reported for only four datasets excluding the VOC dataset, due to the unavailability of its saliency results for the VOC dataset. The decrease in threshold from 255 to 0 (to compute the precision and recall values) corresponds to the increase in recall values from 0 to 1. As it can be observed that at  $T_f = 0$  when all the pixels are retained after segmentation (considered as foreground), all the methods have precision values between 0.2 to 0.4. This indicates that on average 20% to 40% pixels belong to the annotated salient regions for all datasets. At the other extreme, the precision values at the minimum recall values for the proposed technique are higher on average than the other methods (with pronounced increase on the difficult BSD and VOC datasets), depicting smoother saliency maps and uniform saliency saliency assignment inside object contours achieved by maintaining more



true positives at higher thresholds. The PR curve for CS [45] shows a rapid descent at minimum recall values. This suggests that the number of correct predictions drops at high thresholds. A possible reason for this behaviour is that a few background pixels are assigned higher saliency values as compared to the foreground pixels.

The proposed FGSopt approach shows performance improvements on SED1 and VOC datasets as compared with the best performing state-of-the-art methods in terms of AUCPR. The DRFI [68] method exhibits better performance as compared with the proposed FGSopt method on MSRA, SED2 and BSD datasets. However, the proposed method shows equivalent average performance as compared with DRFI (Figure 5.6 (f)). It is hypothesized that the substantial improvements of the proposed FGSopt approach as compared with several state-of-the-art methods on this benchmark are complemented by the optimisation of AUCPR as part of the objective function (5.1) of the proposed FGSopt approach.

### **Segmentation Using Adaptive Thresholding**

Using the adaptive segmentation benchmark, the average precision, recall and F-measure values are reported in Figure 5.7 and the visual results of segmentation on representative images are shown in Figure 5.8. Figure 5.7 (a)-(e) show the precision, recall and F-measure results of methods on individual datasets, while Figure 5.7 (f) shows the average performance of methods on all the datasets. Again for MND [157], average results are reported for four datasets, excluding VOC. The proposed FGSopt method exhibits performance improvements for two of the five datasets on this benchmark. An anticipated reason for this improvement can be attributed to two important artifacts of the saliency maps produced by the proposed method: 1) Due to feature computation on downsampled images, foreground pixels are incorrectly predicted for a few cases, which results, in the loss of foreground pixels during segmentation. A representative example is shown in the first row image of Figure 5.8. 2) For a few specific

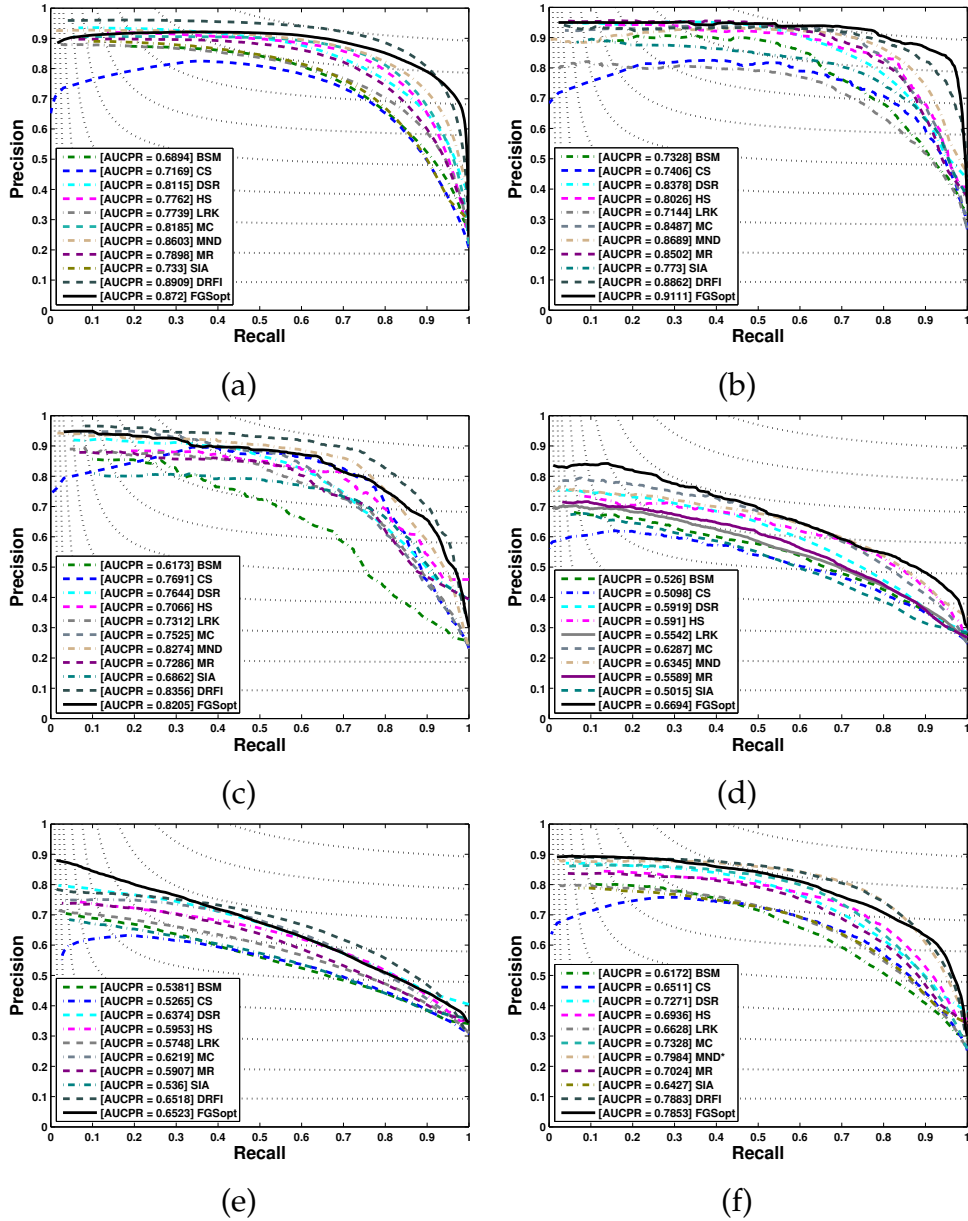


Figure 5.6: PR curves for all methods using fixed thresholding. (a) MSRA, (b) SED1, (c) SED2, (d) SOD, (e) VOC and (f) Combined performance on all datasets. AUCPR for each PR curve is also shown. MND\* in (f) implies that the results reported for MND are average results for only four datasets excluding the VOC dataset. This figure is best viewed in colour.

cases, background noise from low-level features (eigenvectors of  $L$ ) is included in the final saliency. A representative example for this scenario is shown in the second row image of Figure 5.8.

However such cases discussed above are dominated by instances where the proposed FGSopt method produces fine saliency maps that produce accurate segmentation. Representative examples are shown in the third and fourth rows of Figure 5.8 and further are shown in Figure 5.9. The last row of Figure 5.8 shows the robustness of the proposed FGSopt approach to a challenging case of saliency detection.

#### 5.4.4 Interpretation of Results

The evaluations on the fixed thresholding benchmark show that none of the methods outperform all others across all datasets. However, the robustness of the proposed FGSopt approach can be observed by the average results on all datasets in Figure 5.6 (f). We believe that our highly discriminative features (specifically  $f^o$ ) in coupling with the ability to learn to segregate the figure from ground at multiple thresholds, helped the proposed technique to overcome the inherent limitations of the state-of-the-art methods (i.e. non-uniform saliency assignment and inappropriate object annotation). With respect to the average precision, recall and F-measure plots in Figure 5.7, again none of the methods outperform all the other methods on all datasets. Figure 5.7 (f) shows that DRFI [68] and MC [67] exhibit slightly better average performance than the proposed FGSopt method, with our proposed method ranked as the third best on average across all datasets. However, the difference in the performance of the top three ranked methods is minimal as can be seen by the average F-measure values plotted above the methods in Figure 5.7 (f). It is anticipated that the drop in performance of the proposed FGSopt method in terms of adaptive thresholding is due to the artifacts of our saliency map discussed in section 5.4.3.

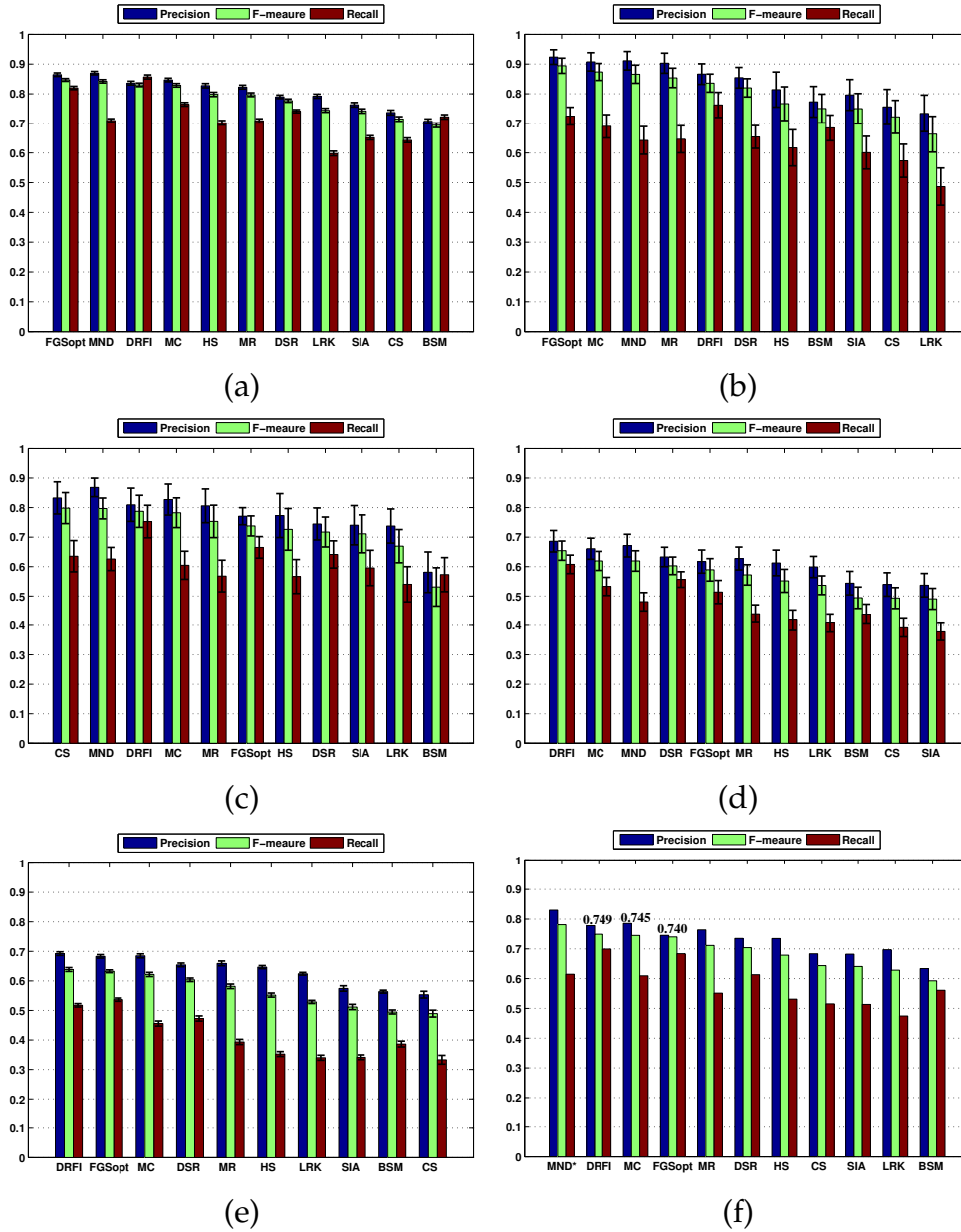


Figure 5.7: Average precision, recall and F-measure for the compared methods on all five benchmark datasets. (a) MSRA, (b) SED1, (c) SED2, (d) SOD, (e) VOC and (f) Combined performance on all datasets. The results are sorted with respect to the F-measure score of the methods. The error bars show the standard deviation from the mean value. MND\* in (f) implies that the results for MND does not include its results for the VOC dataset.

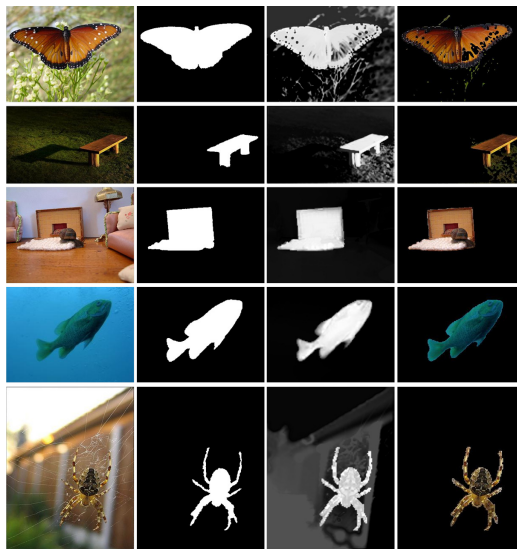


Figure 5.8: From left to right (columns): Input, Ground truth (GT), the proposed saliency maps and the corresponding segmentations performed according to the adaptive thresholding benchmark.

The substantial contribution of this work, which is to overcome the problems of inaccurate object annotation and non-uniform saliency assignment is clearly evident from the visual comparisons of the methods on representative images in Figure 5.9. HS [147] employs local region-based contrast to capture a region's saliency, where each region is an outcome of a watershed-like operation on the input image [147]. As the watershed segmentation does not consider the object boundaries during segmentation and HS [147] only computes the local contrast of regions, the saliency output of HS only manages to capture parts of the salient object as shown in the 3<sup>rd</sup> and the 12<sup>th</sup> rows of Figure 5.9.

MC, MR and DSR are graph based saliency approaches that rely greatly on the boundary prior to compute saliency. These methods capture similarities between regions dominated by the colour cue. Due to over segmentation before saliency computation, these methods struggle in scenarios where the boundary regions share similar colour with the salient object or in cases when a part of the salient object is present at the image border. This results in characteristic responses such as inappropriate annotation or non-uniform saliency assignment to different parts of the salient object. Representative examples of such complex cases for the quoted methods are depicted in rows 3<sup>rd</sup>, 9<sup>th</sup>, 12<sup>th</sup> and 13<sup>th</sup> of the Figure 5.9.

MND [157] captures the perceptual similarity of regions. It mainly employs the colour contrast cue, which is similar to the proposed approach. MND clusters similar regions to identify the distinct regions that are likely to be salient object parts using its global saliency, while it uses a local saliency cue to highlight regions that stand out from their surroundings. The characteristic non-uniform saliency response of MND can be seen in the 6<sup>th</sup>, 8<sup>th</sup> and 12<sup>th</sup> rows of Figure 5.9. For MND global and local cues capture fine aspects of region-level saliency at both scales. However the simplistic feature fusion scheme using super-pixel multiplication affects the final saliency output, where the regions inside the salient object are not properly highlighted as compared with the background regions and

non-uniform saliency assignment can be observed by the representative saliency outputs of MND shown in Figure 5.9.

Heuristic feature selection approaches such as CS and SIA sometimes annotate unwanted background information as salient object due to inappropriate feature selection. LRK learns to separate the foreground information from background based on bottom-up and top-down features, instead of optimally combining the features. This results in inclusion of background noise in scenes with cluttered background.

DRFI [68] performs multiple segmentations of the input image and learns a regressor that directly maps the raw features for each region to a saliency score. The learned regression method is then employed at the test stage to predict the saliency scores for each region and the final saliency is computed by fusing saliency maps computed over multiple segmentations. DRFI exhibits robust performance on multiple datasets and on both benchmarks at the cost of computing an immense 93-dimensional feature vector for each image adding to its computational time (see Table 5.2).

DRFI’s background information features and generic region properties features generally suppress background regions alleviating inappropriate object annotation as can be seen by a few visual examples in Figure 5.9. However, as is common to all region based approaches, the non-uniform saliency response of DRFI can be clearly observed from 5<sup>th</sup>, 9<sup>th</sup>, 10<sup>th</sup> and 12<sup>th</sup> rows of Figure 5.9. This response can be attributed to the regional contrast features of DRFI, which do not respect object composition and assign considerably different saliency to different scales. It is evident from the visual response of DRFI that the properties of proto-objects are not retained in the final saliency, despite inclusion of multiple scales.

The above-mentioned cases show that over-segmentation of images, without respecting important object properties as an initial step in saliency computation occasionally produces unsatisfactory results for saliency methods. Performing object aware segmentations to produce fine-grained object candidates enables the proposed approach to overcome these limita-

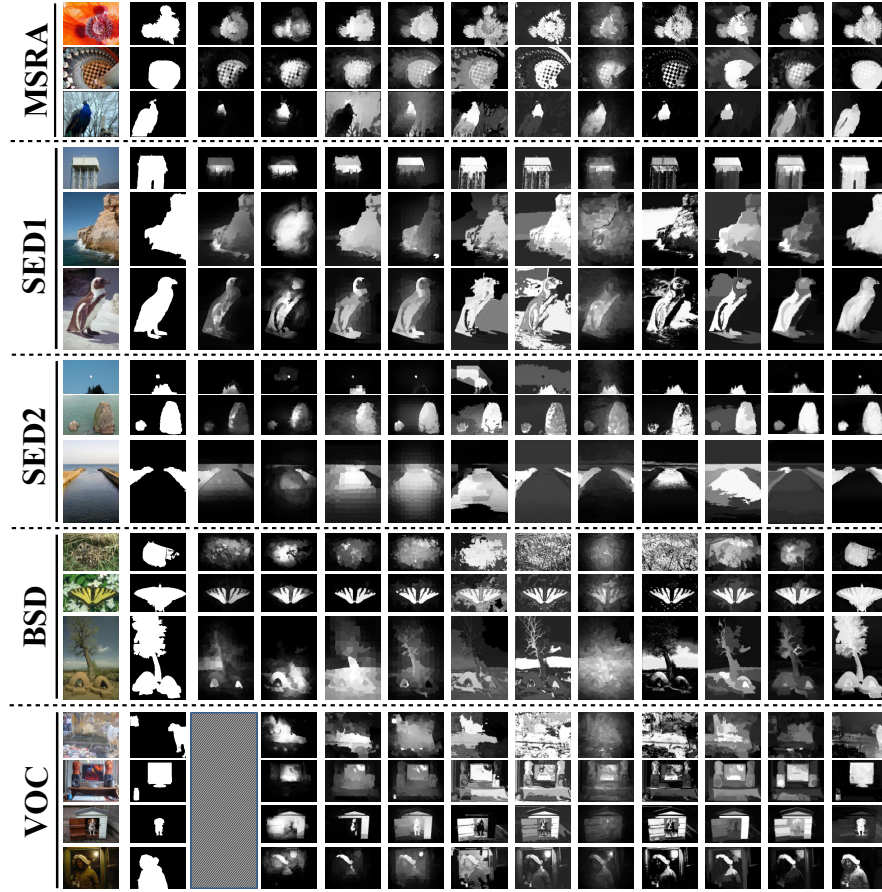


Figure 5.9: Visual comparison with the state-of-the-art approaches. From left to right (columns): Input, ground truth, MND [157], DSR [91], MR [148], MC [67], BSM [144], CS [45], LRK [123], SIA [26], HS [147], DRFI [68] and the proposed FGSopt method. A pattern filled box is shown in column three of Figure 5.9 for the last four rows. This box signifies the unavailability of the saliency maps for MND [157] method for the VOC dataset.



tions. Additionally, optimal combination of features with joint optimisation of important parameters boosts its over all performance.

The timing comparison of the FGSopt method with three learning based benchmark methods is presented in Table 5.2.

Table 5.2: Timing comparison with learning based benchmark methods. The standard deviation results are also included with a  $\pm$  sign.

Method	Time (seconds)
DRFI [68]	13.23 $\pm$ 1.31
LRK [123]	14.50 $\pm$ 1.22
BSM [144]	24.64 $\pm$ 1.62
FGSopt	<b>7.30<math>\pm</math>1.29</b>

The timings reported here were computed using an i7 vpro 3.2 GHz processor with 8 GB of RAM for 100 randomly sampled test images from the MSRA dataset. It can be observed that FGSopt can achieve comparable quantitative performance and better qualitative performance with low computational overhead as compared with other learning based benchmark methods. Several reasons that add to the computational time of the benchmark methods as compared with the proposed FGSopt method are as follows:

1. DRFI includes high dimensional features at multiple scales and evaluates a regressor for each input image.
2. LRK computes 53 dimensional features followed by mean shift segmentation for feature vector representation. Moreover, the low rank matrix (background and the sparse matrix (salient regions) are recovered by Robust PCA, which is the most time consuming step for LRK.
3. BSM employs Laplacian subspace clustering and solves an online

constrained minimisation problem for every image, adding to its computational time.

In contrast, the proposed FGSopt computes nine dimensional features and perform weighted combination of features to obtain the final saliency. As the weights are learned offline, the computational overhead at test stage is low.

## 5.5 Chapter Summary

This chapter introduced a framework for fusion of the FGS output with the smallest eigenvector and the CSD features, by joint optimisation of the feature parameters and feature importance.

The proposed FGSopt method elevated the performance of the FGS method by combining it with complementary features and taking into account the performance gaps amongst individual methods. Quantitative improvements of 3.8% were obtained (in terms of AUCPR) by the proposed FGSopt over the FGS method by taking aid from the eigenvectors and the CSD features. The confidence limits on average AUCPR confirmed that FGSopt improved upon the performance of the FGS method for all the runs.

In terms of comparison of the proposed FGSopt with the baseline un-optimised method, substantial improvements were obtained at feature level. The reason for this substantial improvement can be partially attributed to the improved features, while the major attributable reason behind this performance is the feature importance weights of the FGSopt method in comparison to the uniformly weighted features employed by the baseline method. The proposed FGSopt method was shown to outperform several state-of-the-art methods on benchmark datasets. Specifically, it showed performance improvements of 1.36%, 4.85%, 5.5% and 2.33% over the best performing state-of-the-art methods in terms of AUCPR on the MSRA, SED1, SOD and VOC datasets, respectively.

This chapter and chapter 3 focused on learning a single set of weights to determine feature importance. The learned single set of weights was employed to weigh the features for all test images. The use of a single set of feature weights limit the generalisation capability of the proposed method to unseen image types. As different features might be important for different images, it is anticipated that learning multiple feature weights depending upon the feature composition of images could lead to better generalisation to unseen images. Additionally, suitable feature conditioning and integration approaches (see chapter 1 and 2) were not explored in conjunction with learning feature importance in this chapter and chapter 3. Therefore, the next two chapters will focus on learning multiple feature importance rules along with suitable feature conditioning and integration approaches for an image with specific feature composition.

It is noteworthy that a random forest regressor may also prove useful for feature combination. An advantage of a random forest regressor is its flexibility to encode a solution using multiple decision trees. However, as all the decision trees are averaged to obtain the final solution in a random forest regressor, learning of multiple combination schemes, each suited to an image type and autonomous selection of a scheme (suited to an image type) can not be formulated in a random forest regressor framework.



## **Chapter 6**

# **Genetic Algorithm Based Feature Combination for Autonomously Identified Image Types for Salient Object Detection**

### **6.1 Introduction**

The previous two chapters were focused on learning various design choices of the computational model of visual attention including important feature related parameters and feature importance. A single set of solutions was learned given the training set. This set of learned weights was then applied to unseen test images to measure the generalisation performance of the saliency detection methods. For a single learned solution to achieve good generalisation on unseen data, an important underlying assumption made by the previously proposed methods is that the images in the test set must not have a high degree of variation in comparison to the images in the training set. In practice this assumption does not always hold, due to the highly varying nature of real world scenes, consequential variance in

the performance of features on an image-by-image basis and rarely when the data splitting results in unbalanced train and test sets. To handle such variations in feature performance on image-by-image basis and improve the generalisation of the method on unseen data, learning of multiple feature combination schemes<sup>1</sup>, each suited to a particular image type, is proposed in this chapter. To identify the type of an image and group images having a common type, a semi-autonomous grouping strategy is introduced, based on a feature composition oriented nearest neighbour search in the feature space. A semi-autonomous grouping is employed for placing images into groups for training the proposed method. At the test stage, the proposed method is capable of selecting an appropriate combination scheme by autonomous identification of image type. The images types are defined explicitly by the feature composition of images. For example, the images having cluttered background will presumably have features highlighting background noise and will constitute a unique type in this paradigm. Hence images could be categorised into various types such as cluttered background, multiple salient objects, one salient object, simple background, large and small salient objects etc.

The goal of the work in this chapter is to compare the generalisation performance of a single combination scheme with the performance of multiple learned schemes (each suited to a particular image type) on unseen image data. The following objectives are identified to achieve this goal:

1. To design a baseline genetic algorithm (GA) that is capable of searching for the optimal combination of feature importance, normalisation and integration schemes.
2. To devise a method for semi-autonomous grouping of images that belong to the same image type. Here semi-autonomous means that the placement of images into groups involves no human interven-

---

<sup>1</sup>A feature combination scheme is defined as the concatenation of feature weights, normalisation operation and the integration scheme.

tion, however, the number of images that can be held by a group is fixed.

3. To train multiple baseline GAs using different image and ground truth groups, in order to identify the optimal combination scheme that is best suited to a particular image type.
4. To compare the quantitative and qualitative performance of the proposed method (utilizing multiple learned solutions-one for each image type) with the baseline approach (with single learned solution-one for each image type) on unseen image data.

## 6.2 Rationale for Learning Multiple Feature Combination Schemes Suited to Image Type

It is intuitively desirable to group those images that have features with similar saliency prediction performance. This follows from the observation that a single combination scheme can suffice for multiple feature combinations, when the features show similar performance. On an individual feature level, this essentially means that if a particular feature has good saliency prediction ability on one of the images in a group then it will likely be able to maintain similar performance on all other images in that particular group. However, the problem becomes multifaceted when it is desirable to match the performance of all the features of a particular image to all the features of another image in the group.

A key observation is that the computed saliency features in the proposed scheme explicitly encode information about the foreground and the background. The saliency prediction features are scaled in the range  $[0,1]$ , where a “0” represents definite background, while a “1” represents a definite foreground. A histogram of computed features for a particular image is thus desired to be binary implying that it encodes the foreground and

the background information at its extreme peaks, i.e. “0” and “1”. Therefore, for continuous feature maps, the height of the bins in the middle portion of the histogram is a strong indicator of feature performance and the amount of background noise<sup>2</sup> included in the features. It follows that bin by bin distance between two such histograms should be able to represent their respective performance in terms of capturing the foreground and background regions. Hence, two histograms separated by a greater distance will have a greater difference in their respective saliency detection performance (using the same saliency detection performance measure) as compared with histograms having a smaller distance between them. Therefore, grouping images that have a smaller distance in the feature space is a reasonable approach to group similar image types. The benefit of searching for the best feature combination scheme for a particular group (covering a particular image type) is increased generalisation on unseen images that belong to the same image type.

An illustrative example is shown in Figure 6.1. The images in the groups are plotted along with their respective feature histograms. The histograms are created by concatenating the pixel by pixel values of all the computed features employed in this work (For details of the features, please see section 6.3.1). The saliency detection performance of the feature group is evaluated in terms of mean absolute error (MAE) (for details on MAE, please see 4.2.3) and displayed along with the images. It can be clearly observed that the histograms of the images in image group one share high similarity both at the peaks with specifically low values in the middle region. This indicates good performance as the peaks indicate highlighting of the salient objects and low values in the mid region depicts low background noise. Being a measure of correct background prediction, the MAE reflects good performance for both the images in group one. The histogram of the image in image group two has more mass in the mid-

---

<sup>2</sup>Background noise means pixels belonging to the background are assigned high intensity making them salient



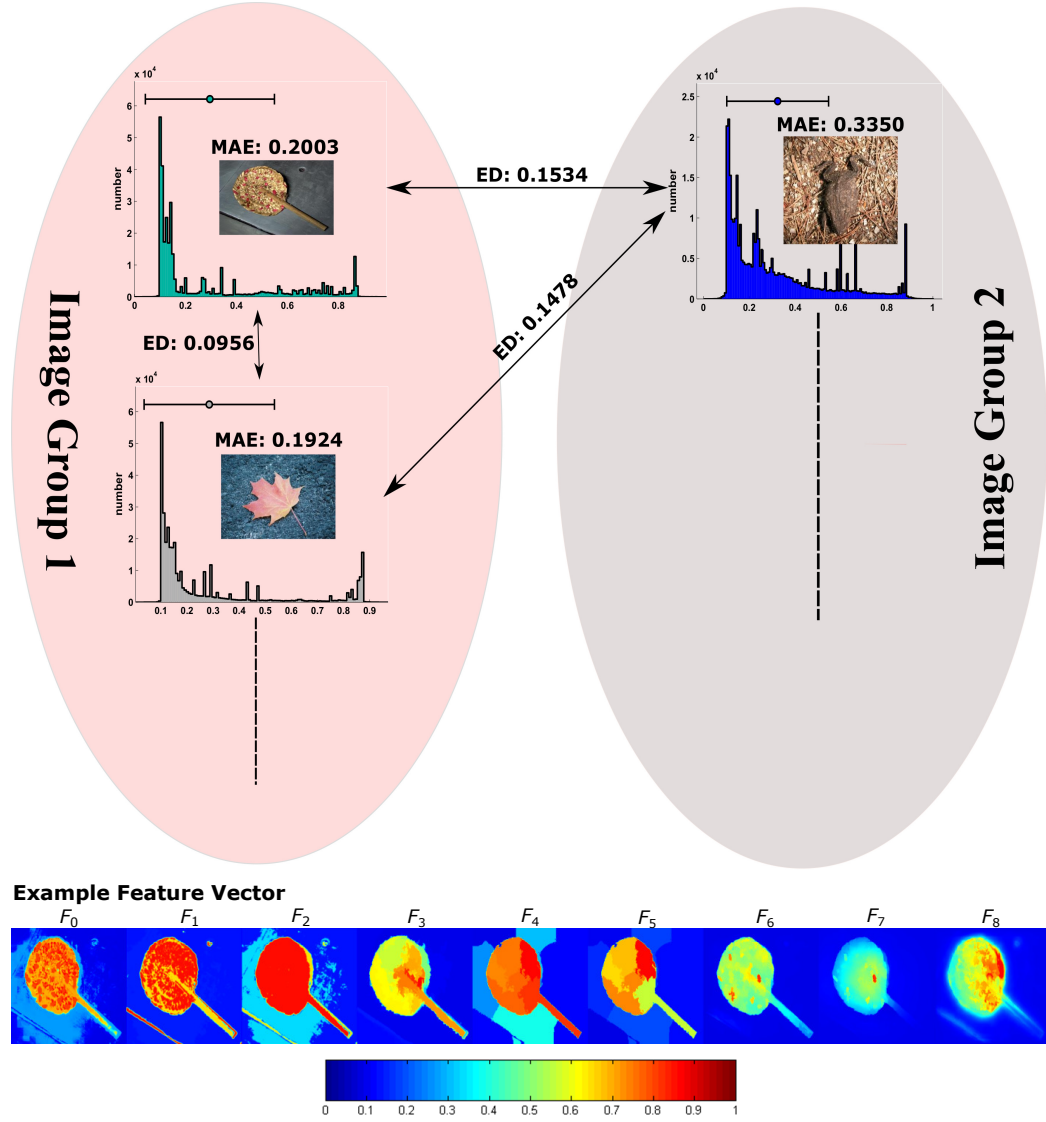


Figure 6.1: Rationale for grouping features before learning feature combination schemes for each image type. The scale depicts the intensity of the features between the range  $[0,1]$ . ED represents the Euclidean distance between the histograms in feature space, while MAE is the mean absolute error of the feature vector after comparison with the ground truth.

dle as compared with the images in group one, which is correspondingly reflected by its higher MAE. The point to be noted is that this difference in performance due to feature composition is accurately reflected by the Euclidean distance (ED) between the features in feature space, where the distance between the image features in image group one is considerably lower than the distances between the images in group one and group two. The Euclidean distance between the histograms is computed as bin-by-bin standard sum of differences. A simplistic distance measure is employed to estimate how the histograms differ specifically in their middle sections, in order to segregate good and bad performing features. Our experiments demonstrate that a bin-by-bin Euclidean distance measure performs well in grouping nearest neighbours based on their performance and a standard distance measure such as Kullback-Leibler or Bhattacharya is not likely to have an improved impact on the performance. Chapter 7 investigates a more sophisticated grouping scheme and a more sophisticated distance measure known as Earth Mover's Distance to observe any performance improvements.

An example of the computed feature vector employed in this work for one of the images is shown in Figure 6.1. The features are scaled between [0,1] and plotted in 2D with same size as the image. The intensity shows the magnitude of the value for a pixel and is a reflection of the respective saliency of a pixel, where dark red represents a "1" and blue represents a "0" as also depicted by the colour bar.

### 6.3 Proposed Methods

The overall architecture of the proposed method termed as image dependent Genetic Algorithm (IGA) is shown in Figure 6.2. It consists of three major steps namely, feature extraction, then image grouping, followed by independent GA based optimisation to learn multiple combination schemes. The process of feature extraction will be detailed in section 6.3.1 and is

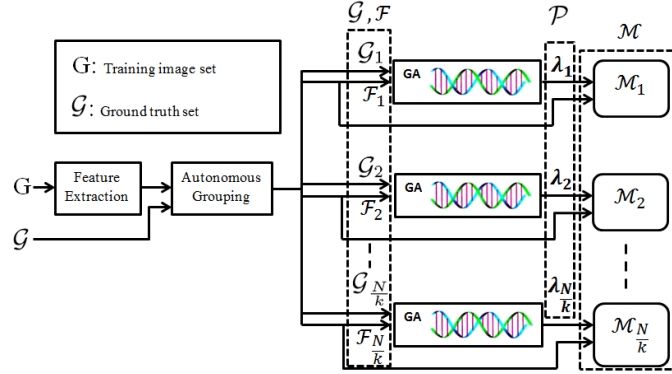


Figure 6.2: System model of IGA. Feature extraction process is defined in section 6.3.1. The process of autonomous grouping is defined in Algorithm 1.

used to extract features for the training image set  $G$ . The procedure for autonomous grouping is depicted in Algorithm 1. The specific details of the proposed Genetic Algorithm are described in 6.3.2. The outputs of multiple GA methods having independent feature and ground truth sets as inputs are connected to multiple independent memory sets  $\{\mathcal{M}_1, \mathcal{M}_1, \dots, \mathcal{M}_{\frac{N}{k}}\}$ . The memory sets hold the learned unique solutions  $\{\lambda_1, \lambda_1, \dots, \lambda_{\frac{N}{k}}\}$  and the corresponding feature groups  $\{\mathcal{F}_1, \mathcal{F}_1, \dots, \mathcal{F}_{\frac{N}{k}}\}$ . For detailed information on the method and the used notation, please refer to section 6.3.3.

### 6.3.1 Feature Extraction

The features that have been demonstrated to correlate with visual attention in prior works and are well-suited for the task of salient object detection are employed for this work. The prime objective of a learner in a feature combination task is to bridge the performance gaps amongst the individual features to enhance the overall combined performance. Therefore, in order to thoroughly evaluate the learning performance of methods,

features having a high degree of variability in performance were chosen such that there is a considerable difference in performance of the lowest performing feature to the highest one.

A total of nine low and mid-level features were extracted from the raw image and are re-sized to 200x200 pixels. The features were resized to reduce the amount of feature data for large datasets. The following features were extracted for each image:

- $F_0$ : One global feature that assigns low saliency to colours that vary a lot in the spatial domain, which is based on the work of Liu et al. [94]. To compute the spatial variance of the colours of an image, all the colours in the image are modelled by Gaussian mixture models (GMM). Afterwards, each pixel in the spatial domain is assigned to a colour component of the GMM. Next, horizontal and vertical variances of each colour component of the GMM are calculated and added to obtain the total variations of colours in the spatial domain. Finally, the colours having high total variance are assigned low saliency, while the colours exhibiting low variance in the spatial domain are assigned high saliency values.
- $F_1$ : One global feature that captures the contrast between clusters obtained through k-means segmentation, inspired by the work of Fu et al. [45]. The contrast cue of a cluster is computed by accumulating its distance to all other clusters, where the distance is actually measured as the distance between cluster centres in feature space using the  $L_2$  norm. The cluster centres are mean pixel features in Lab space.
- $F_2$ : One global feature computing the spatial distribution of pixels in a cluster with respect to the image centre, inspired by the work of Fu et al. [45]. The global spatial distribution of pixels in a cluster from the image centre is measured as the mean of Euclidean distance between pixels in a cluster to the image centre.

- $F_3$ : One region based feature that computes the global contrast between spatial neighbouring regions only [29]. The image is first segmented using graph-based image segmentation. Next, a quantised colour histogram is constructed for each region. Afterwards, saliency for each region is computed as its weighted colour contrast to all other regions in the image. The weights are the number of pixels contained by a region to emphasize contrast to larger regions, while the contrast itself is measured by the colour distance metric between the two regions.
- $F_4$ : One mid-level feature that uses the objectness of image windows to highlight salient objects, based on the work of Alexe et al. [8]. The objectness measure for a window is based on four image cues. The first cue is multi-scale image saliency based on the work of Hou et al. [55]; the second cue is the colour contrast of a window from its surrounding regions, computed as the Chi-square distance between the Lab histograms of the window and its surrounding superpixels; the third cue is computed by measuring the edge density inside a window; finally, the last cue counts the number of superpixels that have their pixels both inside and outside the boundary of the window. The number of such superpixels that have their pixels both inside and outside the window boundary must be low for a window having a high probability of containing the object(s).
- $F_5$ : One feature that groups regions based on their objectness score. Similar regions in terms of objectness scores are merged to form a larger region. For each region, its difference from all other regions is computed in terms of objectness scores to form a difference matrix, whose size is equal to the square of the number of regions. From the difference matrix a global threshold is calculated by finding the smallest differences that exist between neighbours. Afterwards, a local process compares regions only with their neighbouring regions

and groups those having a difference less than the global threshold by assigning them the same objectness.

- $F_{6,7}$ : Two low-level region-based colour features adopted from the work of Naqvi et al. [105]. One colour feature for each region is computed by accumulating the earth mover's distance (EMD) of its Lab histogram from the histograms of all other regions in the image, while the other one is computed by measuring EMD between the histograms of a region and its neighbouring regions only.
- $F_8$ : One feature that highlights salient patterns based on the work of Naqvi et al. [105]. The salient patterns are determined by finding any outstanding patches that have a large distance from neighbouring patches. Match distance is employed to compute the histogram distance between patches due to its ability to capture cross-bin similarities/dissimilarities.

### 6.3.2 Genetic Algorithm

Real-coded representation of chromosomes is used as the majority of the decision variables in the parameter vector  $\lambda_i$  (refer to the next subsection for details) are real. The normalization and integration functions are encoded as integers in the Genetic Algorithm (GA), while the feature weights are encoded as real numbers. Roulette wheel selection is employed, which imposes a trade-off between convergence rate and the quality of individuals retained. As the focus of this work is to improve quantitative performance improvement with low emphasis on the computational time requirements, roulette wheel selection is employed to increase the chances of fitter solutions to stay in the optimisation process and improve the quality of final solutions. Uniform crossover and a custom mutation function is used to compute the next generation (for details please refer to the next subsection). The iterations of the GA solely depend upon the number of generations set by the user.

**Example Pattern**

Weights Vector	Normalization	Integration
$w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9$	$x, \exp(x), -\frac{1}{\log(x)}, \frac{1}{1 + \exp(-x)}, x(M - \bar{m}), x + x * DoG$	$\sum_{i=1}^n x, \prod_{i=1}^n x, \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{x}\right)^{-1}$

**Specific Example**

$-0.01, 0.92, 0.07, 0.41, 0.02, 0.23, 0.99, 0.10, -0.11$	$x + x * DoG$	$\sum_{i=1}^n x$
--	---------------	------------------

Figure 6.3: A pattern of the phenotype and specific example of an individual in the proposed GA population. Possible parameters and functions that can be encoded are shown by the example pattern.

**Encoding Decision Variables**

The chromosomes use real-coded representation with integer constraints. A pattern of the phenotype and a specific example of an individual is depicted in Figure 6.3.

The weighting parameters for the features are encoded in the range  $[-1, 1]$ . Each weight is set as in (6.1).

$$w \in [-1, 1]. \quad (6.1)$$

Six different normalization schemes are encoded as follows. Each integer represents a different normalization operation.

$$\mathcal{N} \in [1 \dots 6]. \quad (6.2)$$

The three integration schemes used in this work are encoded in (6.3). Each integer represents a single integration function.

$$\odot \in [1 \dots 3]. \quad (6.3)$$

The  $i^{\text{th}}$  optimal parameter vector to be searched, denoted by  $\lambda_i$  can be obtained by concatenating the weight vector  $w$ , normalization operation  $\mathcal{N}$  and the integration function  $\odot$ . Section 6.3.3 explains the methods used to learn multiple such parameter vectors that are employed in this work.

### Objective Function

The problem of saliency learning is modelled as a binary classification problem, as in previously reported methods [13, 154]. The goal of the objective function is to maximize the classification accuracy of the method. To achieve this goal, the objective function is set as to minimize the difference between the ideal classification accuracy and the computed classification accuracy. In order to compute the classification accuracy of a particular saliency map, the saliency map output for the method is first computed as follows:

$$S = \odot_i^n w_i \mathcal{N} F_i. \quad (6.4)$$

Afterwards, the saliency map is segmented using a threshold. Based on prior works, a suitable value for the threshold is empirically searched as a multiple of the mean pixel value of the saliency map [55, 3]. The segmented saliency map is then compared with a binary ground truth map and the number of correct (TP and TN) and incorrect (FP and FN) predictions are accumulated to compute the classification accuracy. The fitness is expressed as

$$O(\lambda_i) = 1 - \frac{TP + TN}{TP + FP + TN + FN}. \quad (6.5)$$

### Crossover

During the initial experiments, both blend crossover [38] and uniform crossover [47] operators were tested. While the blend crossover thoroughly mixes the parents' traits, it is easier to encode and enforce integer constraints in the uniform crossover. Moreover, the uniform crossover function was found to converge to better solutions in our experiments. To this end, the constraints arising from the bounds on decision variables are incorporated.



### Mutation

To traverse the search space and find optimal solutions, it is desirable to mutate our individuals, where a function is randomly replaced by another function or a weighting parameter is replaced by a random weight. Unlike crossover, the next solution is searched by adaptation based on the last successful or unsuccessful generation. The random search directions and step size take into account the bounds on variables. This is achieved by ensuring that the mutated solutions generated after the incorporation of search vectors are within the bounds defined on the variables. A combination of step size  $s$ , scale  $sc$  and a direction vector  $\mathbf{u}$  are added to the parent chromosome  $\mathbf{p}$  to compute the offspring. This procedure is depicted as

$$\mathbf{o} = \mathbf{p} + s \times sc \times \mathbf{u}. \quad (6.6)$$

The variable  $sc$  defines the scaling of the variables. Direction vector  $\mathbf{u}$  and the step size  $s$  account for the search directions, keeping the search within the bounds on variables. The step size is adapted based on the ratio of successful generations to the total number of generations. The step size determines the probability of mutating a variable and the size of changes for each mutated variable. The direction vector takes into account the step size and computes the specific variations that are to be added or subtracted from each individual gene of a mutated solution.

#### 6.3.3 Image Dependent GA Based Approach

**Notation :** The training image set is denoted as  $G = \{G_1, G_2, \dots, G_{\frac{N}{k}}\}$ , where the  $i^{\text{th}}$  image group is represented by  $G_i \subseteq G$ . The complete feature set for the training images is denoted by  $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_{\frac{N}{k}}\}$ .  $\mathcal{F} \in \mathbb{R}^{M \times N}$ , where  $M$  is a product of  $D^2$  and  $n$ .  $D$  is the dimension of a single feature ( $D=200$  here),  $n$  is the number of features ( $n=9$  here) and  $N$  is the number of images in the dataset. The  $i^{\text{th}}$  feature set for the  $i^{\text{th}}$  image group

$G_i$  is denoted as  $\mathcal{F}_i \subseteq \mathcal{F}$  and is given as  $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_k\}$ , where  $i \in [1 \cdot \frac{N}{k}]$ .  $\mathbf{f}_i \in \mathbb{R}^M$  is a feature vector and  $k$  is the number of nearest neighbours.

The complete ground truth set for the training set  $G$  is denoted by  $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_{\frac{N}{k}}\}$ . The  $i^{\text{th}}$  ground truth set for the  $i^{\text{th}}$  image group is denoted by  $\mathcal{G}_i \subseteq \mathcal{G}$ ,  $i \in [1 \cdot \frac{N}{k}]$ . The optimal parameter set is denoted by  $\mathcal{P} = \{\lambda_1, \lambda_2, \dots, \lambda_{\frac{N}{k}}\}$ , where  $\lambda_i \subseteq \mathcal{P}$  is the  $i^{\text{th}}$  optimal parameter vector for the  $i^{\text{th}}$  image group. The  $i^{\text{th}}$  memory set represented as  $\mathcal{M}_i \subseteq \mathcal{M}$  is comprised of a feature set  $\mathcal{F}_i$  and a parameter vector  $\lambda_i$ . Here  $\mathcal{M}$  is the set of all memory sets  $\mathcal{M}_i$ . For a particular image in the test phase, the optimal parameter vector is found by searching for the closest image group in the feature space and is denoted by  $\lambda^*$ .

The procedure for the training process of the IGA is depicted in Algorithm 1. The images are autonomously placed into groups depending upon their feature composition. The process starts by searching for  $k$  nearest neighbours for an image based on its distance to other images in feature space. The metric used for the  $k$  nearest neighbour search employed in this work is the Euclidean distance. The image features and ground truth along with its nearest neighbours are assigned to the current groups  $\mathcal{F}_i$  and  $\mathcal{G}_i$  and deleted from the complete feature set  $\mathcal{F}$  and the ground truth set  $\mathcal{G}$ , respectively. This process is repeated until the number of features in the complete feature set  $\mathcal{F}$  falls below the nearest neighbours  $k$ . After the images are divided into groups, multiple GA methods are trained, each with corresponding features and ground truth from different groups. The resulting optimal parameter vector  $\lambda_i$  obtained from an independent GA and the corresponding feature set  $\mathcal{F}_i$  for the  $i^{\text{th}}$  group is stored in the corresponding memory set  $\mathcal{M}_i$ .

During the testing phase of the IGA, the feature vector  $\mathbf{f}_t$  for each test image is computed. For the  $i^{\text{th}}$  feature set  $\mathcal{F}_i$ , the sum of Euclidean distances (denoted by  $d$ ), between its feature vectors  $\mathbf{f}_i$  and the test feature

---

**Algorithm 1:** Training Process of IGA
 

---

**Data:**  $\mathcal{F}, \mathcal{G}$ **Result:**  $\mathcal{M}$ **// Sub Procedure for Autonomous Grouping****while**  $|\mathcal{F}| \geq k$  **do**    Find  $k$  nearest neighbours for the first element of the set  $\mathcal{F}$ ;    Assign the feature vectors for the current element and nearest neighbours to the set  $\mathcal{F}_i$ ;    Assign the ground truth for the current element and nearest neighbours to the set  $\mathcal{G}_i$ ;     $\mathcal{F} \setminus \mathcal{F}_i, \mathcal{G} \setminus \mathcal{G}_i$ ;     $i++$ ;**// Sub Procedure for Training Multiple Genetic Algorithms****for**  $i \leftarrow 1$  **to**  $|\mathcal{F}_i|$  **do**

Train each GA according to the settings described in section 6.3.2;

    Find the optimal parameter vector  $\lambda_i$  for each independent GA;    Create a memory set  $\mathcal{M}_i = \{\lambda_i, \mathcal{F}_i\}$  for each independent GA;

vector  $\mathbf{f}_i$  are computed according to

$$d = \sum_{i=1}^{jF_{ij}} \|\mathbf{f}_i - \mathbf{f}_t\|. \quad (6.7)$$

The sum of distances for all groups are concatenated to form a vector  $\mathbf{d} = [d_1, d_2, \dots, d_{\frac{N}{k}}]$ . The shortest distance  $D$  between the test vector and a feature set is computed as  $D = \min(\mathbf{d})$ .  $D$  is then used to access the corresponding memory set and the optimal parameter vector  $\lambda^*$ . Afterwards, the saliency is computed and thresholded to yield a binary saliency map. The binarised saliency is employed to compute the precision and recall at the test stage. This process is depicted in Algorithm 2.

---

**Algorithm 2:** Testing Process of IGA

---

**Data:**  $\mathcal{M}$

**Result:**  $\hat{\text{Prec}}, \hat{\text{Recall}}$

Compute feature vector  $\mathbf{f}_t$ ;

**for**  $i \leftarrow 1$  **to**  $|\mathcal{F}_i|$  **do**

Compute  $D$  as discussed above;

Use  $D$  to find the corresponding  $\lambda_i = \lambda^*$ ;

Compute saliency using the learned parameters according to (6.4);

Segment the saliency map and compute  $\hat{\text{Prec}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$ ;

Compute  $\hat{\text{Recall}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ ;

---

## 6.4 Design of Experiments

### 6.4.1 Datasets and Experimental Setup

The feature comparison based grouping strategy of the multiple combination schemes presented in this work can be effectively tested on two

classes of images: 1) images where the foreground and background regions share similar features 2) scenes with multiple salient objects. Therefore, this work uses real world images from the challenging Weizmann segmentation evaluation database (SED) [9]. The database contains 200 images (of varying aspect ratio) in total, divided into two datasets (100 images for each dataset), namely SED1 and SED2. The two datasets include one and two foreground objects respectively. Unlike other databases, SED includes images with cluttered background and multiple salient objects, resulting in increased difficulty of segmentation.

The pixel wise ground truth segmentations are also provided by the SED database [9]. The segmentations contain two classes for one object images and three classes for the two object images. The ground truth segmentations have the same dimensions as the input images and are obtained by manual segmentations by three different human subjects. The individual annotations are processed to obtain the ground truth segmentations according to the method prescribed by [9]. The binary ground truth segmentations for both two and three class images are acquired by thresholding the votes from each human subject for each foreground pixel. If there exists at least two votes for a particular foreground pixel, then it is assigned “one” and the remaining pixels are assigned “zero”.

To evaluate the robustness of a saliency map, the standard benchmark method that is reported in [3] and is known as fixed or naive thresholding is employed. According to fixed thresholding, the continuous saliency map is thresholded at a fixed threshold  $T_f$  within  $[0, 255]$ . To compare the saliency response from different methods, this threshold is varied from 0 to 255 to produce a precision-recall (PR) curve. In order to induce accurate segmentations, the continuous saliency map must be able to preserve a high precision as well as a high recall. Therefore an F-measure curve is also plotted by varying the thresholds in the same range, i.e.  $[0, 255]$  in order to measure the salient object segmentation quality of the methods. F-measure is computed using Equation 5.8 as discussed in the previous

chapter. The F-measure curves of the methods reflect the quality of the salient object segmentation capability of their saliency maps.

The images for both SED1 and SED2 datasets are randomly split into train and test sets using the ratio reported in past works [22, 9]. A number of values were experimented with to find suitable values for the population size and the number of generations for the GA. Initial experiments were conducted with a large population size of 1000 and with high number of generations, i.e. 200. However, further experimentation revealed that reducing the population size to 200 and the number of generations to 50 yields similar solutions. After ensuring that the best solutions are retained in every generation by keeping six elite individuals, the crossover fraction is set high with a value of 0.8 with the expectation of finding better individuals.

#### 6.4.2 Selected methods for Comparison

In relation to the goals of this chapter, the GA method discussed in section 6.3.2 was selected as the baseline method in comparison with the proposed IGA method. The IGA method employs multiple GA methods to learn multiple combination schemes and has the ability to select a suitable combination scheme depending upon image type at the test stage.

In addition, two benchmark learning methods, namely Linear Support Vector Machine (SVM) [72] and Non-linear SVM [13] were included for comparison due to their state-of-the-art performance and learning ability for achieving generalisation on unseen images. The major difference in the proposed approach and these benchmark methods proposed in prior works [72, 13] is that they only learn feature weights. Therefore, following the prior works [72, 13], features are integrated by simple linear summation without any normalisation. Learning of the most suitable normalisation and integration schemes is not easy to formulate in the framework used by the SVMs, so was not included in the test. The implementations

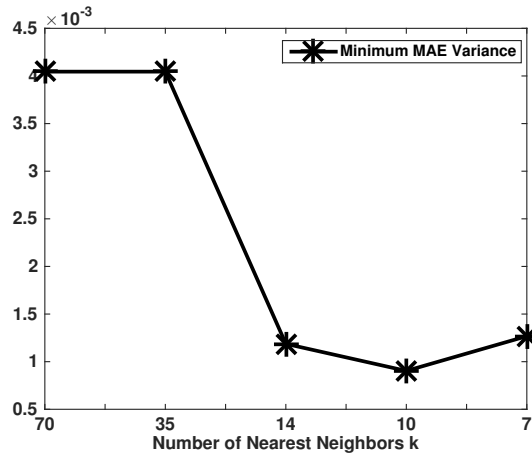


Figure 6.4: Effect of the number of nearest neighbours  $k$  on grouping performance.

details for these benchmark methods are the same as reported in section 4.3.4.

Three deterministic state-of-the-art methods namely AC [2], FTS [3] and MSSS [5] are also included for a more thorough comparison of the proposed approach. The methods are selected as they are specialised for the task of salient object detection.

### 6.4.3 Training Experiments

The baseline GA method is trained using the training sets for both the SED1 and SED2 datasets in order to learn distinct solutions per dataset. Cross-validation is not employed, as standard data splitting is used [22, 9]. To find a single representative solution from 30 independent runs of the GA for each dataset, the Euclidean distance between each of the 30 solutions and mean solution is searched and the solution that has the minimum distance is selected as the final representative solution.

The most important parameter of the IGA method is the number of nearest neighbours  $k$ . To determine a suitable value for  $k$ , a quantita-

tive measure for evaluation of a grouping is required. Considering the rationale behind grouping, a group is considered to be good if the images contained by it have highly similar salient object detection performance. Therefore, the problem of finding a suitable value for  $k$  is formulated as finding  $k$  that results in minimum mean absolute error (MAE)<sup>3</sup> variance across groups by sweeping  $k$  for various values. The minimum MAE variance is obtained by finding the variance for each group and selecting the minimum. Figure 6.4 plots the minimum MAE variance for a range of values for  $k$ . It can be clearly observed that there is a decrease in minimum MAE variance until  $k$  reaches 10, where it starts to increase. It is also worth noting that decreasing  $k$  results in more image groups, which in turns require more independent GA's to be trained. With this in mind, the number of nearest neighbours  $k$  is set to 10 for the proposed method. This choice of  $k$  results in seven image groups for the training phase. For each training image group, an optimal solution is searched using the baseline GA implementation. Each training experiment for a single GA run takes approximately 2.7 hours with the GA settings defined previously. The training time for the GA is considerably higher than the SVM based benchmark methods due to the evolutionary process. However, both the GA and SVM based techniques take similar time at the test stage for a single image.

## 6.5 Results and Discussion

### 6.5.1 Comparison of IGA with the Baseline GA

In this section, the performance of the baseline GA approach is compared with the proposed IGA method in terms of average PR and F-measure curves, plotted in Figure 6.5. The results for the baseline GA are plotted in black, while the results of the proposed IGA method are shown in red.

---

<sup>3</sup>For details of MAE, please see section 4.2.2.



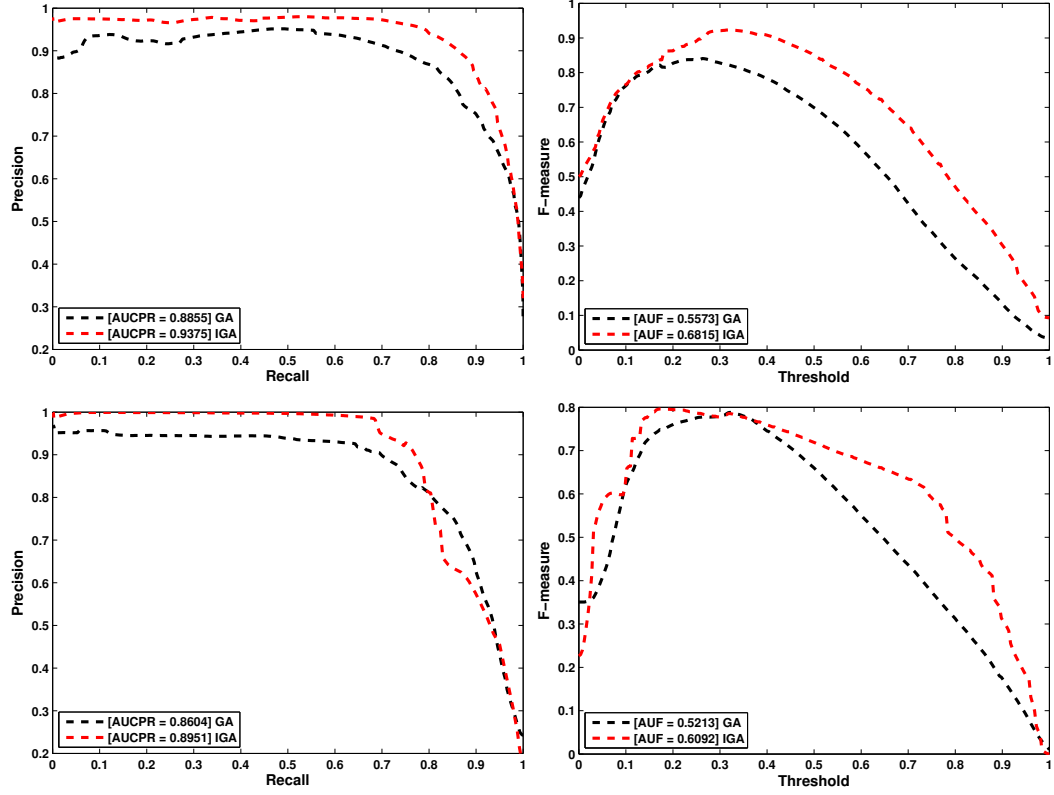


Figure 6.5: Quantitative comparison of the baseline GA method with the proposed IGA method. Top row: average PR curve and average F-measure curve on the SED1 test set. Bottom row: same set of results for the SED2 test set.

The proposed IGA method obtains moderate improvements in terms of AUCPR and substantial improvements in terms of AUF over the baseline GA method on the SED1 and SED2 test sets. From the PR curves, it can be observed that the proposed IGA method maintains high precision values against very low recall values. This result suggests that the saliency maps produced by the proposed IGA approach are smoother and assign higher saliency inside object contours as compared to the baseline GA.

From the discussion in section 6.4.1, it follows that the F-measure curves signify the quality of induced segmentation of a saliency map. Therefore,

the substantial improvements achieved by the IGA in terms of AUF support the hypothesis of superior segmentation quality of the saliency maps produced by IGA. This result is confirmed by the visual saliency results in Figure 6.7 and specifically by their corresponding salient object segmentations in Figure 6.8.

### 6.5.2 Comparison of IGA with Existing Work

This section presents a quantitative comparison of the proposed IGA method with learning based benchmark methods namely, LSVM [72] and NLSVM [13], and selected state-of-the-art methods. Figure 6.6 plots the performance of the methods in terms of PR and F-measure curves for the SED1 and SED2 test sets.

The proposed IGA method achieves considerable improvements in terms of AUCPR and immense performance gains in terms of AUF as compared to the best performing state-of-the-art methods on the SED1 and SED2 test sets. It is worth noting that the performance enhancements obtained by the proposed approach in terms of AUF are even higher than the performance gains achieved over the baseline GA. This result signifies that despite achieving reasonable performance on the fixed thresholding benchmark, the compared methods do not perform well on the task of salient object segmentation as reiterated by their respective segmentation outputs in Figure 6.8. Possible reasons for this poor saliency prediction behaviour of the compared state-of-the-art methods are discussed in the next section.

LSVM and NLSVM exhibit similar performance in terms of both AUCPR and AUF measures on both test sets. This is in accordance with our initial experiments to find suitable parameters for SVM with linear and non-linear kernels.

Pairwise two-sided rank sum test were performed to investigate the statistical significance (if any) of the proposed IGA method in comparison with the other methods. Table 6.1 shows the p-values for pairwise com-

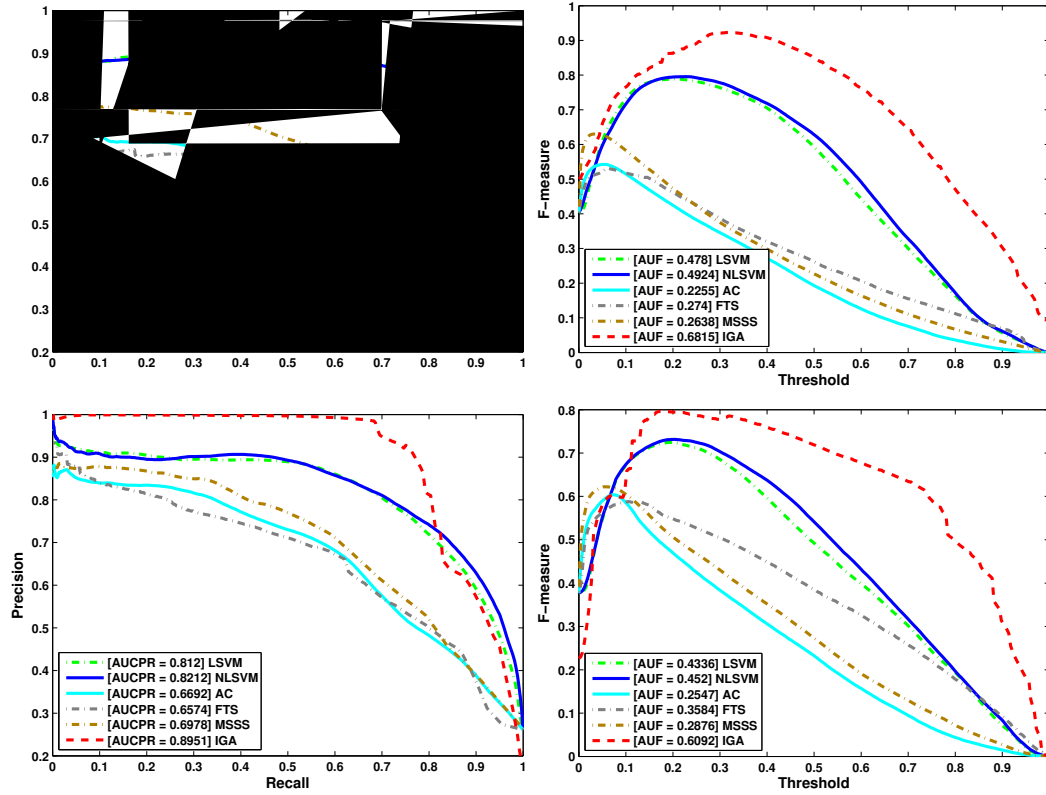


Figure 6.6: Quantitative comparison of the proposed IGA method with learning based benchmark methods and state-of-the-art methods. Top row: Average PR and F-measure curve on the SED1 dataset. Bottom row: Similar set of results for the SED2 dataset. These figures depict the performance of all the methods only on the test images.

Table 6.1: Statistical comparisons of IGA with other methods. The first row shows p-values for AUCPR comparison, while the second row presents the values for AUF comparison.

	GA	LSVM	NLSVM	AC	FTS	MSSS
<b>AUCPR</b>	0.005	0.0027	0.0001	4.69E-08	9.06E-08	2.68E-06
<b>AUF</b>	0.0003	0.0015	0.0004	5.97E-09	2.38E-07	1.73E-07

parison of the IGA method with other methods. It can be clearly observed that the IGA method is statistically different from benchmark methods (i.e. GA, LSVM and NLSVM) and a high statistical difference is observed as compared with the deterministic methods (AC, FTS and MSSS) at the 0.01 level.

### 6.5.3 Qualitative Comparison

Figure 6.7 shows the visual comparison of the saliency output of several methods. The deterministic methods especially FTS [3] and MSSS [5], perform well by assigning low saliency to background but struggle in assigning high values inside salient object contours. As the methods are based on filtering, they can filter out important salient information that lies in the same band with the background information. LSVM and NLSVM mostly highlight object boundaries and can not appropriately weigh features that score higher inside salient object contours. The baseline GA method performs better than other state-of-the-art methods, but misses important salient information in some cases. The proposed IGA method produces the best saliency maps as compared with other methods, highlighting the salient object and suppressing the background. For example, it effectively captures the neck and head region in third row image of Figure 6.7, which is effectively missed by all other approaches.

### 6.5.4 Analysis of Evolved Solutions

Table 6.2 shows the set of representative solutions learned by the proposed IGA method. The first three rows of Table 6.2 present the learned solutions for three representative image groups from the SED1 dataset, while the last three rows present the learned solutions for three representative image groups from the SED2 dataset, respectively. Table 6.3, shows the representative solutions learned by the baseline GA method for SED1 and SED2 datasets, respectively.

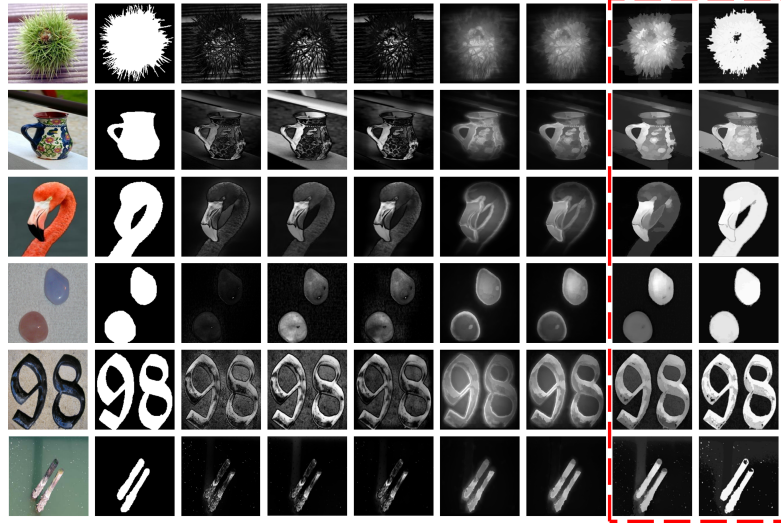


Figure 6.7: Visual comparison of methods on representative test images from both datasets. From left to right: Input, GT, AC, FTS, MSSS, NLSVM, LSVM, GA, IGA. Our GA and IGA methods are shown in the red box.

Table 6.2: Representative learned solutions by the proposed IGA method for representative image groups from both SED1 and SED2 datasets.

	Parameter Set										
	$w_0$	$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	$w_6$	$w_7$	$w_8$	$\mathcal{N}$	$\circ$
<b>SED1</b>	0.14	0.26	0.50	0.82	-0.26	0.98	0.30	0.30	0.67	5	3
	0.15	-0.42	0.07	0.29	0.07	0.79	0.82	0.22	0.31	1	1
	-0.18	0.00	-0.05	0.08	0.32	0.34	0.06	0.97	-0.01	1	1
<b>SED2</b>	0.12	0.07	-0.01	0.98	0.30	-0.01	0.52	0.03	0.17	6	1
	0.48	0.38	-0.03	0.95	0.02	0.27	0.45	0.07	0.01	1	1
	0.12	0.07	-0.01	0.98	0.30	-0.01	0.52	0.03	0.17	6	1



Figure 6.8: Salient object segmentation induced by saliency methods on representative test images from the SED1 and SED2 datasets. From left to right: Input, AC, FTS, MSSS, NLSVM, LSVM, GA and the proposed IGA. The baseline GA method and the proposed IGA methods are shown by the red box.

Table 6.3: Representative learned solutions for the SED1 and SED2 datasets by the baseline GA approach.

	Parameter Set										
	$w_0$	$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	$w_6$	$w_7$	$w_8$	$\mathcal{N}$	$\circ$
<b>SED1</b>	0.03	0.01	-0.01	0.95	0.91	0.90	0.98	0.36	-0.01	1	1
<b>SED2</b>	-0.01	0.92	0.07	0.41	0.02	0.23	0.99	0.10	-0.11	6	1

For the SED1 dataset, the baseline GA method finds no normalisation in the final solution. This can be explained by considering the simple nature of most SED1 images, where no feature conditioning is generally required. This result is consistent with the learned normalisation for two of the representative image groups by the proposed IGA method. However, the proposed IGA method found global normalisation to be effective on one of the groups. Also for one of the groups in SED1, the effective normalisation scheme was found to be iterative similar to two groups for the SED2 dataset. Iterative normalisation means multiple filtering operations of a map with a difference of Gaussian filter. The anticipated reason for this result is that a few images in the SED1 dataset and quite a few belonging to the SED2 dataset result in features that contain a high degree of falsely highlighted background. This result can be attributed to the difficult background for a few SED1 images, and to the presence of multiple salient objects in addition to difficult background for the SED2 images.

The favoured integration type was found to be linear summation for both the baseline GA and the proposed IGA method. This result will negatively affect the generalisation performance of the baseline GA method as inappropriately conditioned and weighted features can corrupt the final saliency output. However, in case of IGA, appropriate conditioning and weighting promotes complementary features, thus making linear summation a suitable choice. Element-wise multiplication of complementary features would promote robust suppression of background noise but at the

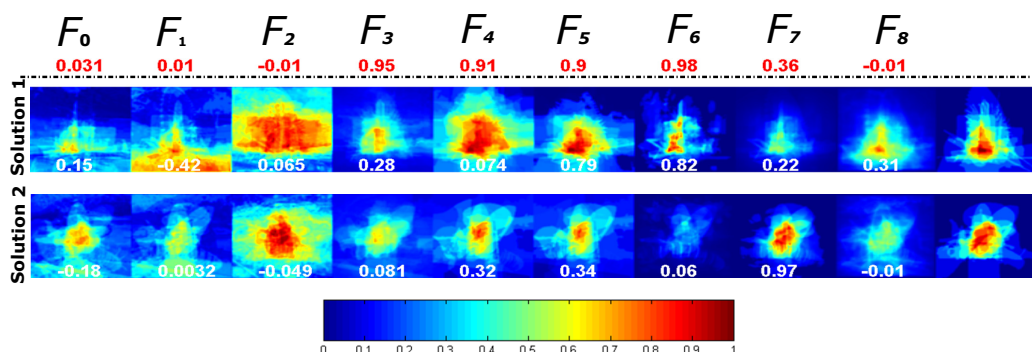


Figure 6.9: Explanation of multiple learned weights by the proposed IGA method. Feature numbers and the learned weights of the baseline GA method for the SED1 dataset are shown above the dotted line. The mean feature maps for the two groups are shown below the dotted line with the learned feature weights overlayed on top. Solutions 1 and 2 signify that the weights are two independent solutions of the IGA method.

cost of substantially decreasing the number of true positives. The harmonic mean was found to be a suitable choice for only one of the image groups indicating that most of the outlier features were removed by appropriate feature conditioning.

In order to elaborate the importance of multiple learned feature weights, a total of 10 images were randomly selected from the SED1 test set and corresponding autonomously identified feature combination schemes were sought. It turns out that the selected images were covered by two of the solutions that were learned from the SED1 dataset by the proposed IGA method. The images favouring the two solutions are placed into two distinct groups covered by distinct solutions. Figure 6.9 plots the mean feature maps for each of the groups with their autonomously identified solutions (feature weights) overlayed on top. The feature weights learned by the baseline GA method for the SED1 dataset are shown above the dotted line. The last column of Figure 6.9 shows the mean ground truth map computed using the ground truth maps of the selected test images.



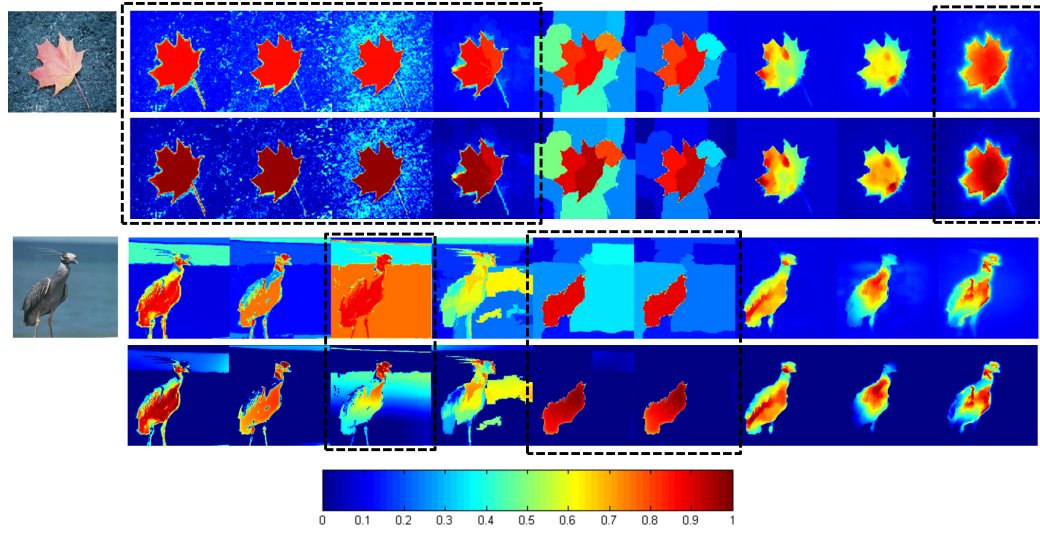


Figure 6.10: Effects of the learned normalisation scheme on representative images.

It can be clearly observed from Figure 6.9, that the weights learned for different feature groups help in generalisation to different image types. The mean feature maps for  $F_5$  and  $F_6$  show maximum overlap with the corresponding mean ground truth map in the first row of Figure 6.9. The good performance of features  $F_5$  and  $F_6$  (for these selected scenes) is rightly captured by solution 1 as it assigns high importance to these feature maps. Similarly, undesirable saliency detection results of features  $F_1$ ,  $F_2$  and  $F_4$  are assigned low or negative importance by the learned solution 1 of the proposed approach. It is to be noted that the mean feature maps for feature  $F_6$  that are covered by solution 2, exhibit high disagreement with the corresponding mean ground truth, in contrast to its performance for the images covered by solution 1. This high variation in performance is truly captured by the low importance assigned to  $F_6$  by solution 2. A similar contrast is found between the weights assigned to feature  $F_7$  by solution 1 and solution 2, respectively. The above examples confirm that the weights assigned to features for different groups in the proposed IGA method help in natural generalisation to unseen images. It can be easily

observed by Figure 6.9 that unlike the proposed IGA, the single feature importance solution learned by the baseline GA can not adapt to the variation exhibited by the unseen test images.

Figure 6.10, shows the effects of normalisation on two representative images from the SED1 dataset. For the two images, the first row shows the actual features, while the second row shows the feature results after global and iterative normalisation, respectively. It can be clearly observed from Figure 6.10 that unlike no normalisation in the first row, the global normalisation results in the second row highlight the salient object (i.e. the leaf) with higher intensity and also suppress the background noise. Important results are highlighted by dashed boxes. Similarly, for the bird image, the iterative normalisation scheme aid in neglecting a high proportion of false background, for a few features highlighted by the dashed boxes.

## 6.6 Chapter Summary

The goal of this work was to compare the saliency detection performance of the single optimal combination scheme learned from all the training images with performance of the proposed multiple combination schemes, where each combination is learned for a specific image type. To achieve this goal, an autonomous image grouping technique was introduced that groups images of the same type based upon their Euclidean distance in the feature space. Subsequently, the image groups were used to train a multiple GA based optimization framework that searches multiple optimal solutions, while each learned solution is particularly suited to a particular image type. The multiple combination schemes and their autonomous identification based upon the image type, helped the proposed method to achieve better quantitative and qualitative performance on unseen images as compared with the baseline single combination scheme. The important findings of this chapter are as follows:

The proposed multiple scheme based method exhibited notable im-

provements in terms of AUCPR and substantial improvements in terms of AUF over the baseline GA approach on unseen test images from the SED1 and SED2 datasets, respectively. The substantial improvements achieved by the proposed IGA method in terms of AUF were shown to be due to superior segmentation quality of the saliency maps produced by the proposed IGA method.

It was shown through analysis of the optimal solutions of the proposed method that the feature combinations learned by the proposed approach adjust according to image types of the unseen test instances, thereby achieving better generalisation as compared with the fixed baseline optimal combination scheme, which fails to adjust according to the varying nature of unseen images.

This chapter proposed a semi-autonomous solution for grouping images to learn multiple combination schemes. The constraint imposed by the number of nearest neighbours in the proposed framework for grouping images could lead to scenarios where two uncorrelated images are erroneously placed in the same group under the same image type. For fully autonomous placement of images into groups, a relative encoding scheme with adaptive condition matching on an image-by-image basis is required. As one possible approach to avoid these problems, the next chapter explores the adaptive classifier conditions of a Learning Classifier System for autonomous matching and placement of input images into groups. Moreover, as Learning Classifier Systems (LCSs) are reported to be more competitive in niches as compared with the traditional GAs, the next chapter explores the potential of an LCS in searching for multiple feature importance rules in comparison with the proposed GA based solution of this chapter.



## Chapter 7

# Learning Classifier Systems Based Multiple Feature Combination Schemes for Salient Object Detection

### 7.1 Introduction

The previous chapter introduced the IGA method for learning multiple feature combination schemes, each suited to a particular image type. The IGA method employed a semi-autonomous scheme for grouping images and the placement of images in a group was limited by the number of nearest neighbours  $k$ . The number of nearest neighbours  $k$  was determined *a priori* for the grouping process based on empirical evidence. Also the histogram distance based scheme compared images based on the collective performance of all the features (feature composition) and did not take into account individual feature level performance variations. To account for the variable group size and also consider the individual feature level variations, a flexible system is required for image grouping through nat-

ural division of the search space. The flexible system should be able to simultaneously cater for feature level variations by matching individual features to autonomously identified image categories. Learning classifier systems (LCSs) have the inherent capability of dividing the search space into niches, complemented with their competitive search ability inside the niches. Additionally, the classifier condition part of the LCSs provides a natural approach to feature level input matching. Therefore, an LCS is proposed in this work to learn multiple feature importance rules.

The goal of this chapter is to compare the generalisation performance of the fully autonomous image grouping scheme with the previously proposed semi-autonomous image grouping scheme on unseen images. It is anticipated that the fully autonomous image grouping scheme will consider individual feature level variations in contrast to the semi-autonomous image grouping scheme proposed in the previous chapter. The following objectives are formulated to achieve this goal:

1. To introduce a relative encoding scheme in the condition portion of the classifiers that is able to match the input at feature level and autonomously place images into groups.
2. To compare and analyse the performance of the fully and semi-autonomous image grouping schemes in terms of grouping similar images.
3. Quantitative and qualitative performance comparison of the proposed LCS method with the IGA method for salient object detection.

## 7.2 Salient Object Detection Using Learning Classifier Systems

For a balanced comparison with the previously discussed IGA method, a supervised approach, similar to XCSCA [84] is adopted in this work. For the supervised approach, the target saliency map (i.e. the ground truth) is

also provided along with the input image during training. The goal of LCS is to evolve saliency maps for each input image such that the error between the computed saliency map and the target ground truth is minimal. The error term used in this work is given in Equation 7.1:

$$\epsilon = 1 - \frac{TP + TN}{TP + FP + TN + FN} , \quad (7.1)$$

where TP and TN are the number of correctly classified positive and negative pixel values between the computed saliency map and the ground truth. FP and FN represents the falsely classified pixels. The learnt rules can then be applied to assign appropriate importance to feature maps for previously unseen test images.

Instead of matching the input image at pixel-level against conditions of classifier rules, nine saliency-based features are computed for each input image, where each feature is a two-dimensional real-valued matrix. The same features as employed by the IGA method, are used in this work for a fair comparison with the IGA method. These features are adapted due to their high degree of variability in performance, in order to thoroughly evaluate the learning performance of methods. For details of the nine features, please refer to chapter 6.

If a classifier's conditions in an LCS are matched directly to the computed image features, then it will be hard to evolve any generalised classifier rules due to the large-sized two-dimensional image features [83]. Therefore, in order to enable generalisation, a novel relative encoding scheme is introduced to match an input image against classifier conditions. In this encoding scheme, each computed image feature  $F_i$  will be encoded as a real-valued variable  $d_i$  to be matched against the classifiers' conditions. The real-valued variable for a feature is computed using earth mover's distance (EMD) [118] from a two-dimensional reference artificial feature. EMD is defined as the minimum cost for transforming one histogram into the other such that there exists a ground distance between the features.<sup>1</sup>

---

<sup>1</sup>The fast implementation of EMD with thresholded ground distances based on the

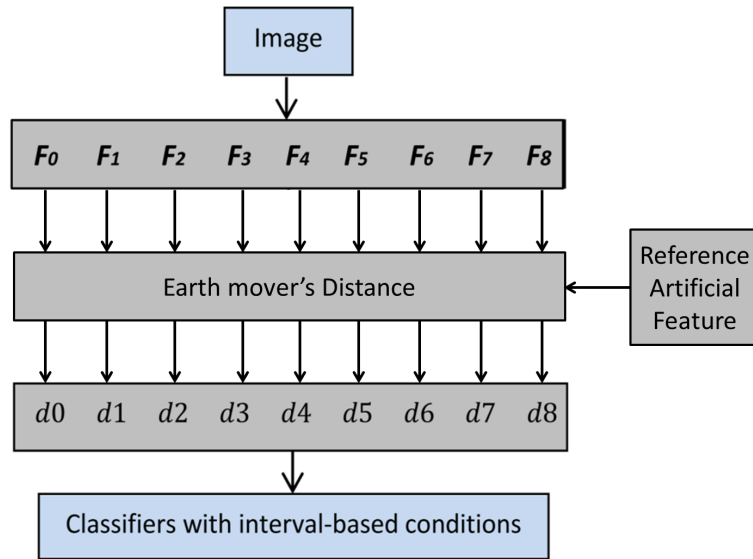


Figure 7.1: The novel encoding scheme to match an input image against the classifier population in order to enable generalisation in classifier rules.

The reference artificial feature consists of all "1s". As the real-valued features are in the range  $[0,1]$ , the ground distance between features and the reference feature acts as a relative measure of inter-feature distance in feature space. A relative performance measure of every feature is recorded in contrast to the histogram based distance strategy employed in the IGA method (please refer to chapter 5). Consequently, conditions in classifier rules will be encoded as a concatenation of real-valued intervals so that the method can evolve generalised classifiers, see Figure 7.1. The condition part of the classifiers are composed of intervals with lower and upper bounds on the real-valued variable  $d_i$  to match input features. The conditions are validated such that smaller lower bounds are ensured as compared to upper bounds of the intervals.

Usually an LCS evolves rules having one fixed set of actions. However, in the problem domains having large number of discrete actions and/or

---

work of Pele and Werman [111] is employed in this work.



continuous actions, it is beneficial to *compute* actions instead of *mapping* actions [61, 84]. In this work, a linear combination function is employed to compute the action, i.e. the saliency map, in a classifier rule, similar to XCSCA [84] that used a linear combination function to compute discrete classifiers actions. The linear combination function learns the feature importance by learning the feature weights.

### Computing Saliency using Linear Combination Functions (XCSCA)

A saliency map is computed as a linear combination of the features of the input image and a weight vector  $w$ . In this function, the computed two-dimensional features are linearly combined with the evolved weights to produce the required saliency map for the matched input image, as given in Equation 7.2. Here  $x_0$  denotes a constant input parameter utilised in the update process of recursive least square.

$$S = w_0x_0 + \sum w_{i+1}F_i; \quad i = 0, 1, 2...8. \quad (7.2)$$

Updating classifier weights using recursive least square is shown to be useful for improving the generalisation and convergence of XCS based systems [85]. Therefore, weights are evolved using recursive least square, according to the procedure described by Lanzi et al. [85]. For updating classifier weights, the error function defined by (7.1) is employed.

## 7.3 Experimental Design

### 7.3.1 Data Set

To evaluate the multiple combination schemes on a widely accepted standard benchmark for saliency evaluation, this work employs the MSRA salient object database [94] using the labelled ground truth from [3]. The

MSRA data set is comprised of 25,000 images in total, a common benchmark dataset in the field and represents many types of datasets as it contains many classes of images. The MSRA data set includes ground truth annotations in the form of labelled rectangles from multiple users. These ground-truth annotations classify multiple objects as a single object by annotating them with a single rectangle and also do not cater for pixel-wise accuracy. To remedy these effects, a set of 1000 images were manually segmented by a single user to obtain binary masks [3]. These ground truth masks consider the effect of pixel-wise accuracy and multiple objects (the details about the ground truth database can be found in [3]), and is widely accepted in the field of computer vision as a standard benchmark for saliency evaluation. The data set includes examples having the following classes of images increasing its diversity and making it difficult for techniques that use a single set of learned parameters: (1) images with cluttered background, (2) images with multiple objects, (3) images with large salient objects, (4) images with small salient objects, and (5) images with faces, persons, objects and text.

The standard fixed thresholding benchmark method reported in [3] is used to generate the precision-recall (PR) and F-measure curves. As the performance of saliency methods also depends upon the number of correctly classified (salient/background) pixels as compared with the total number of pixels of an image, a curve for the classification accuracy is also included as a performance measure. The curve for the accuracy is also computed using the fixed thresholding benchmark. The area under the PR, F-measure and classification accuracy curves denoted by AUCPR, AUF and AUACC, respectively, are also reported for quantification of the results.

### 7.3.2 Experimental Setup

A covering classifier is generated by using the following steps: a classifier condition is created and validated, weight vector  $w$  is initialised with zeros, while all other parameters are assigned according to [21]. During classifier update, the classifier error is updated by Equation 7.1, weight vector is updated according to the action function [85], while the classifier fitness is updated from the classifier error and numerosity as in [137]. For rule discovery, the GA works as in [18, 138].

The XCSCA method use the parameter values suggested by Butz and Wilson in [21] that are commonly used in the literature: learning rate  $\beta = 0.2$ ; fitness fall-off rate  $\alpha = 0.1$ ; error threshold  $\epsilon_0 = 10$ ; fitness exponent  $\nu = 5$ ; threshold for GA application in the action set  $\theta_{GA} = 25$ ; two-point crossover with probability  $\chi = 0.8$ ; mutation probability  $\mu = 0.01$ ; experience threshold for classifier deletion  $\theta_{del} = 20$ ; fraction of mean fitness for deletion  $\delta = 0.1$ ; classifier experience threshold for subsumption  $\theta_{sub} = 20$ ; reduction of the fitness  $fitnessReduction = 0.1$ ; parameters for integer conditions  $r_0 = 0.7$ ,  $m_0 = 0.5$ ; and the selection method is tournament selection with tournament size ratio 0.4. GA subsumption is activated whereas action set subsumption is deactivated. The number of classifiers used is 2000 and the number of training instances is 20,000, in all the experiments conducted here. Explore and exploit problem instances are alternated. The constant input parameter  $x_0 = 0.5$  is set at centre of the input features range, i.e. [0,1]. As the initial parametrisation is uncertain, a high value of  $\delta_{rls} = 100$  is employed to ensure a fast initial update rate.

For the IGA method, a suitable value for the most important parameter, i.e. the number of nearest neighbours  $k$ , was empirically determined to be 30. For the GA method employed in the IGA method, the settings described in chapter 5 were found to be suitable and used in all the experiments.

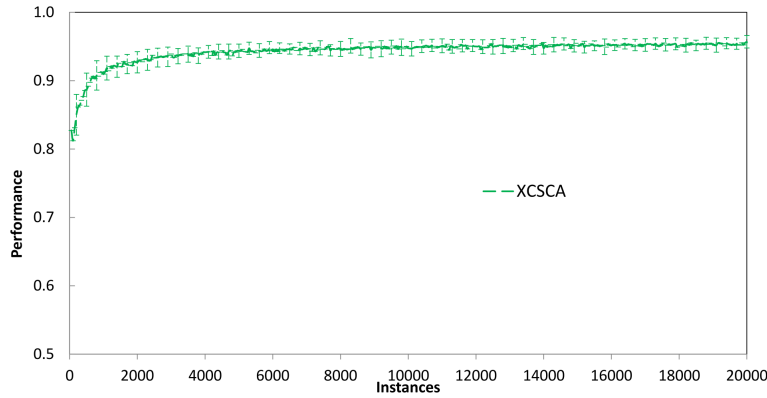


Figure 7.2: Training process of the XCSCA method for detecting salient objects.

## 7.4 Results and Discussions

The performance of the XCSCA method for detecting salient objects in the training process is shown in Figure 7.2. In Figure 7.2, the X-axis is the number of problem instances used as training examples, the Y-axis is the performance measured as the percentage of correct classification during the last 50 exploit problem instances.

The XCSCA method exhibited quick learning in the early stages and reached approximately 96% performance within the first 5000 instances. The error bars show the standard deviation of the 30 runs and confirm the consistency and confidence in performance of the XCSCA method.

The performance comparison of the XCSCA method with the IGA method for detecting salient objects in the testing process is shown in terms of PR, F-measure and classification accuracy curves in Figure 7.3. At recall = 1 in Figure 7.3 (a), when all the pixels are considered as foreground after segmentation, both the methods have a precision of 0.2. This suggests that on average 20% pixels belong to the annotated salient regions for the test set of ASD. At the other extreme of the PR curve, where maximum threshold

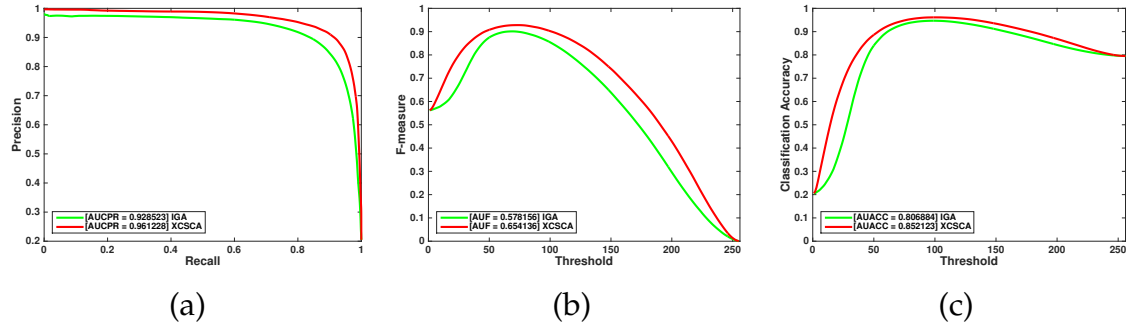


Figure 7.3: Comparison of the XCSCA method with the IGA method in terms of PR, F-measure and accuracy curves. AUCPR, AUF and AUACC results are also reported for comparison.

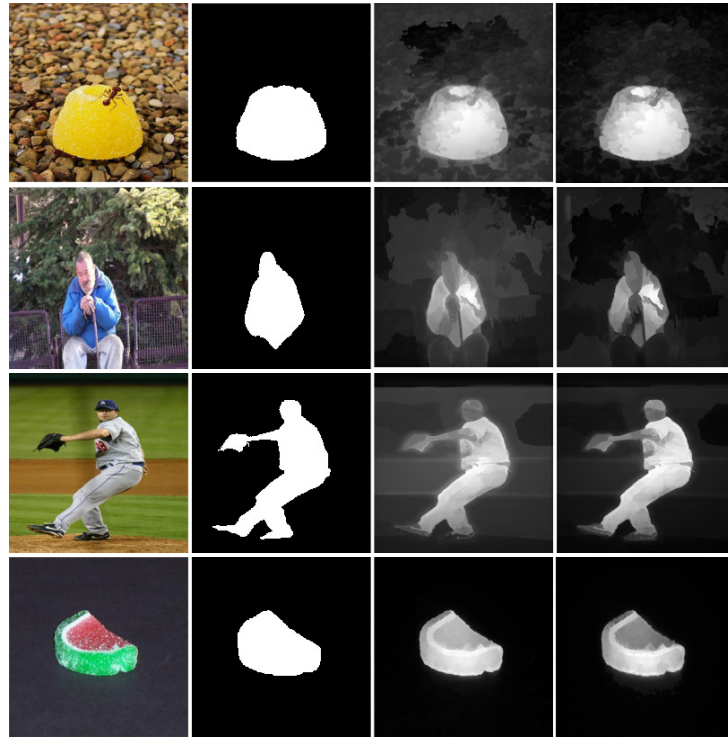


Figure 7.4: Visual comparison of the saliency maps. From left to right: input image, ground truth, IGA and XCSCA.

is applied, the XCSCA method maintains slightly higher precision values than the IGA method. This result suggests that the saliency maps of the XCSCA method are smoother than that of the IGA method inside object contours. Figure 7.3 (b) suggests that the saliency maps produced by the XCSCA method can induce better segmentations than the IGA method. Figure 7.3 (c) shows that the IGA approach falsely highlights some background pixels on lower thresholds as compared with the XCSCA method, which maintains high number of correct predictions.

On average performance improvements of 3.5%, 13% and 5.7% were observed in terms of area under PR, F-measure and classification accuracy curves by the XCSCA method. Results of a two-sided t-test produced p-values of 0.0022, 0.0004 and 0.0014 for the three performance measures demonstrating the statistical significance of the results. The t-test was performed after confirming the normal distribution assumptions.

Figure 7.4 presents the visual comparison of saliency maps generated by the XCSCA and the IGA methods. Higher intensity values mean higher saliency and vice versa. The XCSCA method suppresses the background noise that is visible in the IGA saliency maps for the first three row images. For the last row image with plain background, the IGA method has smoother saliency inside the object contour.

## 7.5 Analysis of Evolved Solutions

Table 7.1 shows three representative solutions learned by the independent GA methods of the IGA method. The independent GA methods were trained using different image and ground truth groups. It can be observed that features  $F_4$  and  $F_8$  are heavily weighted by all the three representative solutions, due to their good saliency detection performance on most images. It is also worth noting that features  $F_6$  and  $F_7$  are generally given low importance with negative weights on a few occasions. This is a reasonable result as these features frequently highlight false background regions.

Notably, a few features that carry a negative weight for one group are assigned a positive weight for another group. This is subject to the feature variability for different image groups. This ability aids the IGA method to assign feature importance depending upon image type.

Table 7.1: Three representative solutions learned by the independent GA methods of the IGA method using unique image groups.

No.	Weight Vector								
	$w_0$	$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	$w_6$	$w_7$	$w_8$
1	0.02	0.20	0.15	0.36	0.85	0.21	-0.40	-0.16	0.98
2	0.30	0.22	0.34	0.60	0.96	0.98	0.19	-0.18	0.73
3	0.06	-0.40	0.28	0.50	0.39	0.32	0.33	0.25	0.97

A sample of the experienced and accurate classifier rules evolved in a typical run of the XCSCA method are shown in Table 7.2. The linear combination based method evolved weights for each feature in each classifier, see Table 7.2. It is to be noted that there are 10 weights in each classifier, the first on the left side corresponding to the constant input  $x_0$  and the remaining nine corresponding to the nine computed saliency features. The classifier conditions are real valued, where each bracket contains two real values separated by a comma. The comma separated real values act as upper and lower bounds on the features' earth movers distance from the reference map and are matched for each input image to decide its covering rule. From Table 7.2, it can be observed that features  $F_4$  and  $F_8$  are again heavily weighted by the XCSCA method similar to IGA. It is also worth noting that features given high importance from one of the rules are also found to be negatively weighted by another rule, thereby adding to the generalisation power of the XCSCA method to unseen images.

Table 7.2: A sample of the experienced and accurate classifier rules, obtained in a typical run, using linear combination based XCSCA method.

No.	Condition	Weight Vector											
1	[0.74,1.00] [0.14,0.25] [0.09,0.72] [0.00,0.82] [0.00,0.56] [0.00,0.49] [0.00,1.00] [0.00,0.68] [0.00,0.98]	35.2	46.3	25.6	-5.5	29.8	26.9	5.7	96.4	1.7	3.3		
2	[0.00,0.74] [0.12,0.64] [0.14,0.59] [0.00,0.70] [0.19,0.36] [0.00,0.56] [0.59,1.00] [0.00,0.02] [0.00,1.00]	-335.6	192.1	181.7	-3.3	369.8	-329.2	691.8	343.7	38.5	733.5		
3	[0.19,1.00] [0.08,1.00] [0.08,1.00] [0.61,0.82] [0.19,1.00] [0.50,1.00] [0.83,1.00] [0.00,0.50] [0.20,1.00]	-799.9	606.5	-138.1	117.5	2.9	441.2	-243.8	200.3	540.2	491.1		
4	[0.19,1.00] [0.74,1.00] [0.62,1.00] [0.50,0.68] [0.35,0.80] [0.00,0.36] [0.41,1.00] [0.00,0.47] [0.54,1.00]	-674.4	594.2	-155.8	150.3	-116.7	560.0	-225.9	124.4	475.7	583.8		
5	[0.18,1.00] [0.12,1.00] [0.00,0.59] [0.42,0.70] [0.00,0.53] [0.00,0.46] [0.59,1.00] [0.18,0.47] [0.25,1.00]	-145.1	-28.5	-1.9	914.7	994.3	-509.5	488.0	-39.4	349.3	-499.8		
6	[0.00,0.74] [0.12,0.51] [0.00,0.59] [0.00,0.68] [0.00,0.60] [0.14,0.46] [0.59,1.00] [0.00,0.02] [0.28,0.34]	40.8	-243.6	-83.5	606.9	532.6	611.4	672.0	69.7	-142.0	174.7		
7	[0.00,1.00] [1.00,1.00] [0.08,1.00] [0.00,1.00] [0.00,1.00] [0.00,0.38] [0.41,0.66] [0.00,0.52] [0.53,1.00]	-629.4	774.4	-697.0	177.8	-8.2	1081.5	660.1	1532.2	101.5	647.9		
8	[0.43,0.83] [0.87,1.00] [0.46,0.97] [0.05,1.00] [0.47,1.00] [0.00,0.02] [0.23,1.00] [0.38,0.90] [0.00,0.31]	-635.9	390.7	-542.1	216.3	-2.9	1077.3	699.7	1604.1	300.7	654.2		
9	[0.43,0.83] [0.87,1.00] [0.00,0.97] [0.05,1.00] [0.47,1.00] [0.00,0.02] [0.23,1.00] [0.00,0.73] [0.00,1.00]	-631.0	391.6	-542.4	214.5	-2.7	1081.2	694.4	1601.3	301.1	656.2		
10	[0.37,0.87] [0.50,0.74] [0.56,1.00] [0.61,1.00] [0.00,1.00] [0.00,1.00] [0.23,1.00] [0.00,0.50] [0.44,1.00]	-931.3	260.1	-342.5	457.2	-3.9	931.7	403.5	1284.7	381.8	730.4		



## 7.6 Analysis of Grouping Schemes

In contrast to the fixed grouping scheme employed by IGA, the XCSCA method groups the images by dividing them into niches. To quantitatively evaluate the grouping obtained by the two approaches, the minimum inter group variance (for all groups) in terms of mean absolute error (MAE), is sought for a varying number of image groups. The rationale behind evaluating grouping quality in terms of MAE variance is that a high quality grouping will place similar images in terms of their MAE performance together in the same group, thus reducing the inter group MAE variance. Figure 7.5 presents a comparison of the XCSCA based grouping and the IGA based grouping strategy. The number of image groups corresponds to the number of solutions in the XCSCA method, while it is a function of the number of nearest neighbours  $k$  in the IGA method. For each number of image groups, varying between 3 and 30, a single value for the variance of the MAE is computed by comparing the inter group variance of all the image groups and selecting the smallest one. As can be seen by Figure 7.5, the minimum inter group variance is dependent upon the number of solutions for the XCSCA method, while on the number of nearest neighbours for the IGA method.

It can be clearly observed in Figure 7.5 that there is a decreasing trend followed by the minimum MAE variance as the number of groups increase. The latter is true for both the compared approaches. This is a reasonable result, as an increase in the number of image groups naturally increases the possibility of more similar images being placed in the same group, thereby decreasing the overall inter group MAE variance. An ideal grouping strategy will form image groups with same MAE performance for all the images in a particular group. It is evident from Figure 7.5 that the XCSCA method groups more similar images into a niche as compared to an image group of the IGA method as shown by the minimum variance of the XCSCA method in terms of MAE.

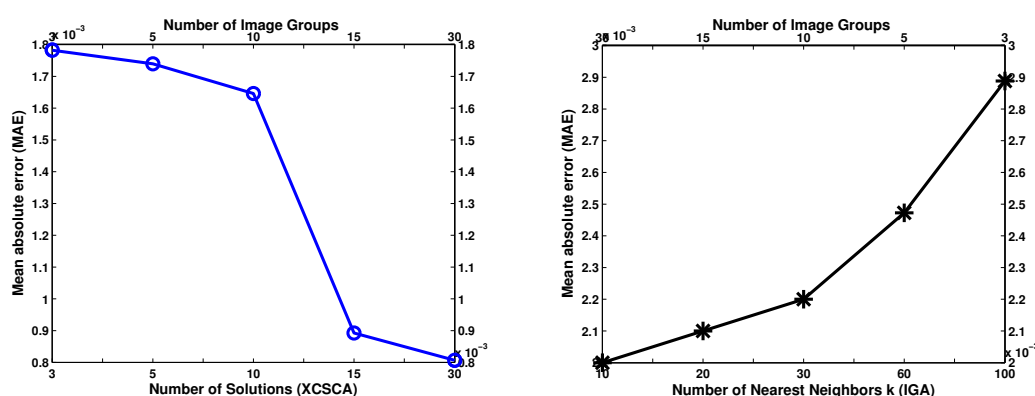


Figure 7.5: Comparison of minimum inter group MAE variance of images between XCSCA and IGA methods. The plot on the left shows the minimum MAE variance with respect to the number of solutions for the XCSCA method. The plot on the right shows variation in the minimum MAE variance as the number of nearest neighbours  $k$  is varied. Note that the scale ranges are different between plots. The number of image groups are displayed on the top of x-axis for both plots.

It is to be noted that there is a steady decrease in the MAE variance for the IGA groupings, while the results for the XCSCA method show a sudden decrease from 10 to 15 groups. A possible reason for this result is that the number of images in a group are steadily decreased in each group for the IGA method, thereby resulting in a steady variance decrease. In contrast there is no constraint in terms of the number of images per group in the XCSCA method and images are adaptively grouped based on their relative performance. Therefore, a sudden decrease in variance is expected depending upon the quality of newly added solutions and the coverage provided by them in terms of capturing similar image types in the same group.

To further explain the image grouping results in Figure 7.5, an illustrative result is presented in Figure 7.6.  $\mathcal{F}_1$  and  $\mathcal{F}_2$  represent a set of feature matrices and the individual features are labelled with feature symbols. The numbers shown below a few features in square brackets are the classifier conditions required to match these features. The features for which no conditions are explicitly shown means that they share the same classifier conditions. The IGA method placed the corresponding images to  $\mathcal{F}_1$  and  $\mathcal{F}_2$  in the same image group based upon the Euclidean distance between their respective histograms as shown in Figure 7.6. Although the overall performance of the two feature matrices is very similar in terms of the extreme peaks as judged by the histogram based approach of the nearest neighbour scheme, the contrast in terms of features  $F_1$ ,  $F_2$  and  $F_8$  is still clearly evident. To differentiate between the potentially different feature matrices, the XCSCA method has two different overlapping rules, where the classifier conditions are the same for most of the features and different for the features that are in contrast to each other, as can be seen from Figure 7.6 conditions.

Although, the XCSCA based grouping method successfully captures the subtle differences in the respective performances of the features  $F_1$ ,  $F_2$  and  $F_8$ , it still struggles to identify the very difficult to detect differ-

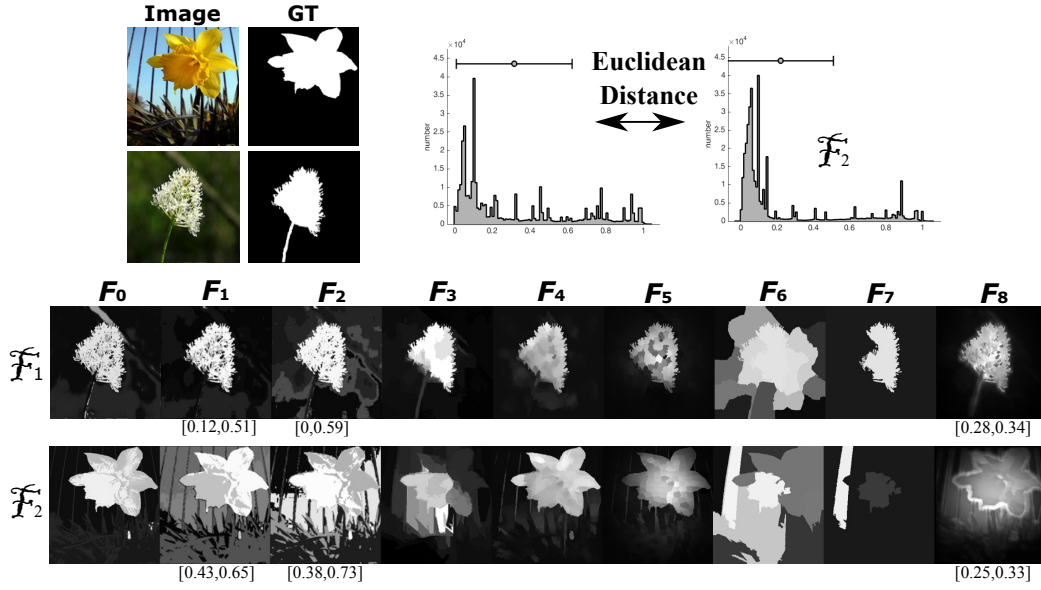


Figure 7.6: Explanation of the intergroup variance results in Figure 7.5.

ence between the features  $F_7$  of the two feature vectors. As can be seen from Figure 7.6, that the feature  $F_7$  of  $\mathcal{F}_1$  captures part of the object and clearly suppresses the background. On the other hand, the feature  $F_7$  of  $\mathcal{F}_2$  falsely highlights the background while completely missing the object. Such types of difference in performance of the features can not be easily captured by the relative condition matching strategy of the XCSCA method. To detect such differences, a technique that either relies on prior knowledge or additional cues is required and is therefore a prospect for future consideration.

## 7.7 Further Discussions

To perform a comprehensive comparison of the performance of both the methods, a categorical comparison of methods is performed by isolating the images on which one of the techniques specifically performs better than the other in terms of AUCPR. On analysis of the categorical results,

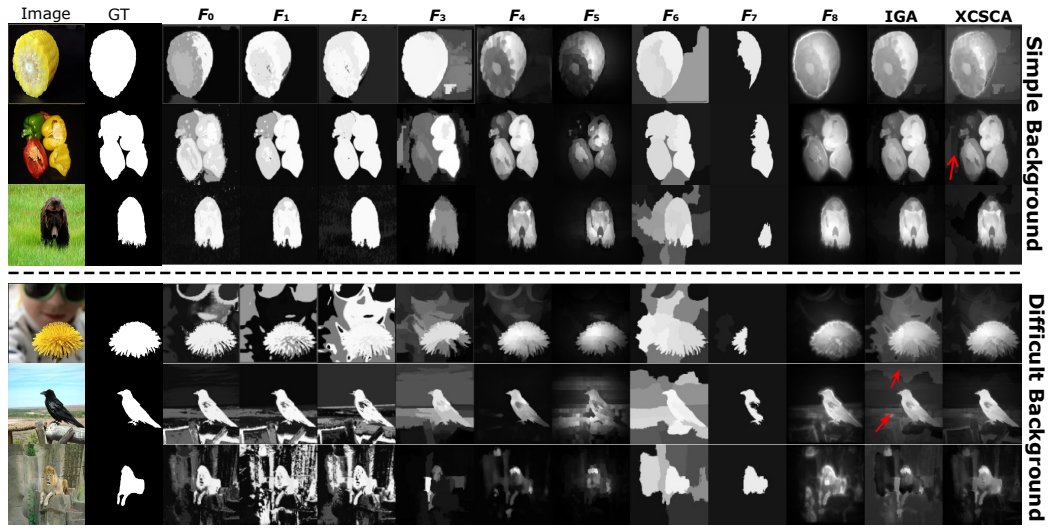


Figure 7.7: Categorical results of the IGA and XCSCA methods on simple and difficult background images.

it was revealed that the IGA method explicitly performs better than the XCSCA method on images with a simple background<sup>2</sup>, while the XCSCA method specifically shows better performance than the IGA method on scenes with difficult backgrounds<sup>3</sup>, in addition to other image classes.

Figure 7.7, presents a comparison of methods on simple and difficult background images. The first three rows of Figure 7.7 shows the comparison on scenes with simple background, while the last three rows present difficult background cases. The intuitive reason behind this result is that, the feature performance on simple background images is highly similar. The absolute histogram based grouping scheme of the IGA method can effectively exploit this fact for simple background images as compared with the relative scheme of the XCSCA method. On the other hand, the relative grouping scheme of XCSCA is more effective in handling varying

<sup>2</sup>Simple background means that it does not contain distractor objects, rich textures or higher colour contrast.

<sup>3</sup>Difficult background implies background containing distractor objects, rich texture or sharing similar colour with the salient object.

performance of features on images with a difficult background, thereby achieving considerably better performance. This observation is clearly explained by Figure 7.7, where the highly similar feature performance for simple background images translates into slightly better overall performance by the IGA method as compared with the XCSCA method both in terms of AUCPR and visual results. For highly varying feature performance for the difficult background images, it can be clearly observed that the relative scheme of the XCSCA based method enable it to achieve considerable performance enhancements over the IGA method both in terms of AUCPR and visual results.

## 7.8 Chapter Summary

The goal of this work was to compare the performance of the proposed fully autonomous grouping of images and learned multiple feature importance rules with the previously proposed semi-autonomous approach for learning multiple rules. To achieve this goal, a supervised XCSCA method was introduced that was able to learn multiple combination schemes based on computation of an action. The XCSCA method was able to autonomously place images into groups and divide the search space into niches for effective learning of multiple importance rules. The quantitative performance of the XCSCA method on unseen images was found to be better than the previously proposed IGA method on the task of salient object detection. The important findings of this chapter are:

Performance improvements of 3.5%, 13% and 5.7% were observed in terms of area under PR, F-measure and classification accuracy curves by the XCSCA method. A two-sided t-test resulted in p-values of 0.0022, 0.0004 and 0.0014 for AUCPR, AUF and AUACC, respectively, demonstrating the statistical significance of the results.

Analysis of the grouping schemes revealed that the images placed in the same group by the XCSCA method, exhibit lower variance of the mean

absolute error (MAE) performance metric, as compared with the MAE variance of grouping obtained by the IGA method. It is anticipated that grouping similar performance images together resulted in more generalised feature importance rules for the XCSCA method. Analysis of the grouping results revealed that the IGA method grouped two features with varying performance on feature level solely based on their distance in feature space. On the other hand, the XCSCA method covered the same images with different rules, where the classifier conditions were found to be same for similar performing features and different for the features that show contrasting performance.

On analysing the visual results obtained from the two approaches, it was revealed that the IGA method shows better performance on images with simple background, while the XCSCA method exhibits notably better saliency response on difficult background images. For detailed performance and timing comparison of the IGA and XCSCA methods, please refer to chapter 9.

This chapter and chapter 6 were focused on finding suitable combination schemes for combining various feature modalities. Multiple combination schemes were proposed to increase the generalisation performance on unseen image types. The next chapter of this thesis investigates how to select the best features for combination and neglect the unnecessary information that negatively affects the final saliency response.





## Chapter 8

# Feature Quality Based Dynamic Feature Selection for Improving Salient Object Detection

### 8.1 Introduction

The last chapter discussed the idea of learning multiple combination rules through autonomous grouping of images using an XCS based method. The motivation was to group similar images based on their feature composition and learn a customised weight vector for each image group. A novel relative encoding scheme was introduced within the XCS framework, which was able to autonomously group images having similar image composition. Although the XCS method was able to identify feature level differences for two feature sets, there was an outlier feature case which could not be detected by the proposed XCS based method. The specific example from chapter 7 is recalled here in Figure 8.1 for discussion. The green boxes show the features successfully identified to be different by the proposed XCS method, while the red box shows the features that could not be differentiated by the relative encoding scheme of the proposed XCS method. If both the features in the red box are assigned similar weights



Figure 8.1: Comparison of image features using XCS's encoding scheme. Green boxes depict the features effectively differentiated by the XCS method, while the red box highlights outlier features, which can not be differentiated by the XCS encoding.

and included in the combination for final saliency output, the bottom row feature will inversely affect the final saliency output, contrary to the first row feature. It follows that discarding the bottom row outlier feature during the combination process will yield better saliency output as compared with the output when it is included in the combination.

The above example urges a need for a system that can identify the goodness or badness of a feature/saliency map, discard outlier features and combine the best performing ones on an image-by-image basis. Once devised, such a system, can be employed within the multiple learner framework proposed in the previous chapter to perform selective feature combination for all image types.

### 8.1.1 Chapter Goals

Following the above discussion, the main goal of this work is to compare the performance of autonomously identified unique feature/saliency map combinations (that are obtained by selecting the best performing features for optimal fusion) with the optimal combination of all feature/saliency maps.

To enable such autonomously identified unique combinations, a dy-

dynamic system is required that can measure the quality of feature/saliency maps in an unsupervised fashion on an image-by-image basis. The desirable system should then be able to select the best performing discriminative feature/saliency maps based on their quality and optimally combine them to obtain the final saliency output on individual image basis. This allows extraneous maps to be neglected.

To realize such dynamic feature selection, the following objectives need to be fulfilled:

1. To devise special feature quality measurement cues that can measure the quality of feature maps in an unsupervised fashion.
2. To devise a clustering method that can cluster good performing and outlier features into distinct clusters.
3. To determine if combining only the best performing features/saliency maps for each image improves upon the performance achieved by combining all the maps.
4. To test the quality measurement ability of the devised features by measuring their ranking ability against ground truth ranking.

### 8.1.2 Chapter Organisation

The rest of this chapter is organised as follows: Section 8.2 briefly reviews the fundamentals of graph based manifold ranking to leverage the understanding of the new graph based formulation for foreground approximation proposed in this work. Section 8.3 explains the implementation details of the dynamic feature selection method proposed in this work. Section 8.4 details the experimental settings. Section 8.5 presents and discusses the obtained results. Finally the last section summarises the important findings of this chapter.

## 8.2 Graph Based Manifold Ranking

A comprehensive review of graph based saliency techniques is presented by [97]. In this section, a brief review of graph-based object saliency is presented and specific details required for better understanding of our new graph-based formulation for foreground prediction are described. All graph-based saliency approaches begin by mapping an image on to a graph  $G = (V, E)$  of  $M$  nodes, where a node  $v_i$  represents the  $i$ th image location (a pixel, patch or a superpixel) and an edge  $e_{ij}$  is the link between nodes  $i$  and  $j$ . Provided initial saliency estimates for all nodes, the goal is to obtain object saliency for all nodes  $\mathbf{o} \in \mathbb{R}^M$  by saliency diffusion along the graph.

The initial saliency estimates (saliency seeds) are denoted by  $\mathbf{s}$ . The edges  $e_{ij}$  are characterized by weights (depending upon the node similarities), which make up the affinity matrix  $\mathbf{W}$ . Saliency seeds are propagated based on the affinities between nodes, which is defined by a propagation matrix  $\mathbf{P}$ . The propagation matrix [37] governs the process of saliency diffusion. A thorough review on the use of various transition matrices for diffusion in the context of image retrieval is presented by [37]. Output saliency nodes  $\mathbf{o}_i$  are assigned saliency values based on their affinities with seed nodes  $\mathbf{s}_i$ . Generally the optimal solution is sought as:  $\mathbf{o} = \mathbf{P}\mathbf{s}$ .

## 8.3 Dynamic Feature Selection (DFS) method

In this section, the proposed method for measuring feature/method quality termed as dynamic feature selection (DFS) method is presented. DFS dynamically selects the features/saliency maps on an image-by-image basis by measuring their quality using high-level cues. The goal is to select only the maps that can boost the overall performance after combination, while discarding the remainder during the fusion process.

### 8.3.1 System Model

The system model for the introduced DFS method is shown in Figure 8.2. Given an image  $I$ , a set of  $m$  feature/saliency maps is computed. Each feature map  $F_i$  is then subjected to two separate operations; foreground approximation (section 8.3.2) and feature quality measurement (section 8.3.3) to compute a vector of feature qualities  $t$ . Afterwards, the best performing features are selected by hierarchical clustering based on feature qualities  $t_i$  (section 8.3.4) to form a set of selected features  $\mathcal{F}_s$ . The selected features from  $\mathcal{F}_s$  are then fused together using the methods described in section 8.3.6 to compute the final saliency output of our DFS method.

The feature quality  $t_i$  for a feature map  $F_i$  is computed using four mid-level cues, i.e., feature density  $f_D$ , sparse reconstruction error  $\epsilon^s$ , feature induced segmentation quality  $s_Q$  and Gestalt contour energy  $e_{\text{contour}}$  (section 8.3.3) with assistance from foreground approximation  $F_G$ .

### 8.3.2 Foreground Approximation

A foreground approximation leading to a robust figure ground proposal is needed to facilitate cues such as feature density, reconstruction error and inter-region contour energy (see section 8.3.3) for measuring feature quality. Most saliency methods rely on a simple contrast of regions to compute saliency. However, it is highly difficult to tackle scenes with a cluttered background and/or multiple salient objects by only utilizing regional contrast. Therefore, such methods can not be directly adapted for this task.

A few object candidate generation approaches such as constrained parametric min-cuts (CPMC) and multiscale combinatorial grouping (MCG) [10, 22] can be adapted to this task. However, these techniques require accurate ranking of object candidates to obtain high quality in terms of object hypothesis. Object ranking is a cumbersome task and efficiency improvements are beyond the scope of this work.

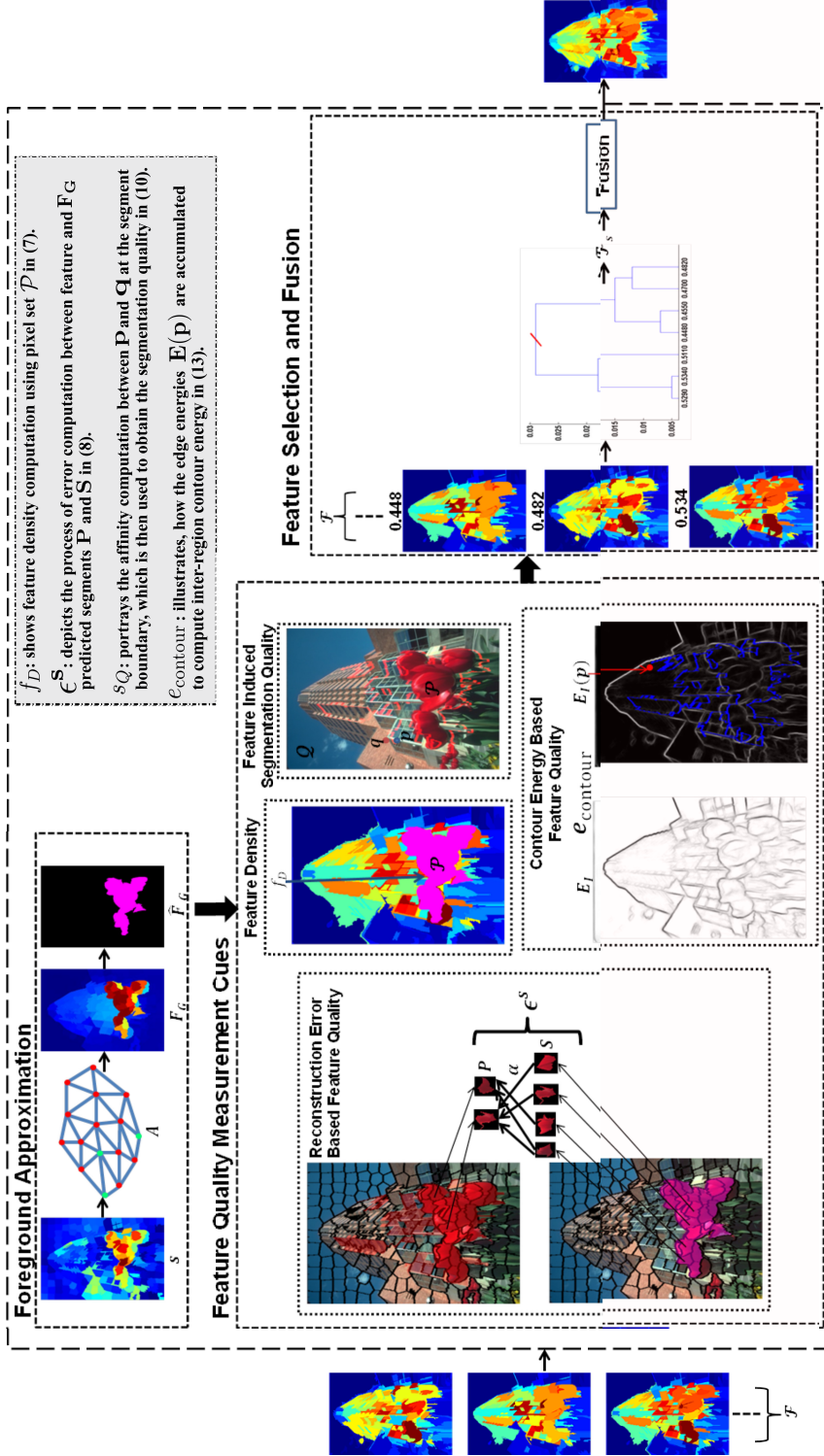


Figure 8.2: System model for dynamic feature selection. The process of foreground approximation, feature quality measurement using the four proposed cues and hierarchical clustering based dynamic feature selection is illustrated. The processes for computing the proposed feature quality measurement cues are also illustrated, where the used symbols are defined at the top right of the figure.

The previous success of graph-based diffusion schemes in approaching the problem of saliency without favouring a certain feature [149, 97] encourages the design of a new graph-based approach for computing the foreground approximation.

## Supapixel Segmentation

To over-segment the input image  $I$ , the simple linear iterative clustering (SLIC) algorithm [4] is employed to obtain  $M$  regions  $x_i$ . Each region (a superpixel) is represented by the mean colour features in the CIE Lab colour space and the image is described as a feature matrix  $[x_1, x_2, \dots, x_M] \in \mathbb{R}^{3 \times M}$ .

## Saliency Seed Prediction

Choosing the right set of seeds  $s$  is a critical task for proper saliency propagation during the diffusion process [97]. Yang et al. [149] used boundary nodes as seeds for ranking. This makes their method vulnerable for images including objects at their boundaries. Seed learning methods such as the recent work of Lu et al. [97] are promising, however high computational time makes it unsuitable for the proposed method. For this work, the regional backgroundness descriptor of Jiang et al. [68] is adapted as opposed to the approach of Yang et al. [149], for initializing seeds in the diffusion process. Although, the backgroundness descriptor also utilise the boundary region to determine seeds, the richness of it high dimensional features and the generalisability of a random forest regressor, makes the backgroundness descriptor a highly discriminative feature for predicting high quality seeds.

The backgroundness descriptor of [68] uses 29 features including average values of colour, histograms of colour, hue, saturation, texton and response of Leung-Malik (LM) filter bank [86], computed region-wise. The LM filter bank consists of first and second derivatives of Gaussian filters

at multiple scales and orientations, constituting 36 filters, 8 Laplacian of Gaussian filters and 4 Gaussian filters. This LM filter bank is highly specialised to detect textures. The backgroundness descriptor feature matrix  $\mathbf{B} \in \mathbb{R}^{M \times 29}$  contains a unique column for each of the features. For the  $i$ th feature the corresponding column vector in  $\mathbf{B}$  is computed as the difference between its feature vector  $\mathbf{f}_i^R$  and the corresponding pseudo-background region feature vector  $\mathbf{f}_i^B$ . The difference for a simple feature is computed as  $\|\mathbf{f}_i^R - \mathbf{f}_i^B\|_2$ . While the difference between a histogram-based feature is computed as  $\chi^2(\mathbf{h}_i^R, \mathbf{h}_i^B)$ , where  $\mathbf{h}_i^R$  and  $\mathbf{h}_i^B$  are histograms of the  $i$ th superpixel region (node) and the  $i$ th pseudo-background region, respectively. The seed vector  $\mathbf{s} \in \mathbb{R}^M$  is then obtained by applying the learned random forest regressor of [68] to the backgroundness descriptor feature matrix  $\mathbf{B}$ . A random forest regressor is an ensemble method that fits a group of decision trees to random sub-samples of the dataset. The output is computed by averaging of decision trees to improve accuracy and prevent over-fitting.

## Graph Construction

Given  $M$  superpixels regions  $\{\mathbf{x}_i\}_{i=1..M}$  as nodes, a single layer graph  $G = (V, E)$  is constructed where  $V$  represents a set of nodes (with each node representing a superpixel) and  $E$  is the set of undirected edges. A  $k$ -regular graph structure is used to exploit the spatial relationship between neighbouring pixels.

Similar to Zhou et al. [155], the affinity matrix  $\mathbf{W} \in \mathbb{R}^{M \times M}$  is constructed by using the following weighting function for two nodes, where  $\boldsymbol{\mu}_i$  and  $\boldsymbol{\mu}_j$  are the mean of superpixels nodes  $i$  and  $j$  in the CIE Lab colour space and  $\sigma$  is a constant used to control the strength of the edge weight

$$w_{ij} = \exp \left( -\frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|}{2\sigma^2} \right). \quad (8.1)$$

The affinity matrix  $\mathbf{W}$  is employed to formulate the propagation matrix



$\mathbf{A} \in \mathbb{R}^{M \times M}$  as

$$\mathbf{A} = (\mathbf{D} - \lambda \mathbf{W})^{-1}, \quad (8.2)$$

where  $\mathbf{D}$  is a diagonal matrix,  $d_{ii} = \sum_j w_{ij}$  and  $\lambda$  is a balancing parameter in the regularization function of (8.2).

Saliency from seeds is then propagated to the foreground approximation vector  $\mathbf{f}_G = \mathbf{A}\mathbf{s}$ . The foreground approximation vector  $\mathbf{f}_G$  is segmented using a threshold of  $\frac{2}{M} \sum_{i=1}^M f_G^{(i)}$  to yield the segmented foreground approximation vector  $\hat{\mathbf{f}}_G$ . The saliency information in each row of  $\mathbf{f}_G$  and  $\hat{\mathbf{f}}_G$  is then assigned to the pixels belonging to the corresponding regions to obtain the original and segmented pixel-level foreground approximation maps  $\mathbf{F}_G$  and  $\hat{\mathbf{F}}_G$ .

### 8.3.3 Cues for Measuring Feature Quality

As the desired response of a salient object detection result is to segment out the complete salient object, a good salient object detection result must ensure two important properties: 1) it must be able to assign uniform saliency inside object contours and cover all parts of the salient object; 2) it must be able to preserve salient object boundaries and ensure high intensity contrast at locations where there is a high feature contrast in the image (mainly the colour feature in this work). Based on these properties, the following feature quality measurement cues are introduced to ensure uniform saliency and high segmentation quality.

#### Feature Density

To induce high quality salient object segmentation, a saliency output must ensure that the average density of the pixels inside the salient region(s) must be considerably higher than the average intensity of any other region of the saliency map. This motivates us to propose a unique cue termed feature density ( $f_D$ ) that measures the quality of a saliency map based on its intensity density inside a predicted foreground region as compared with

its intensity density for the rest of the map. The predicted foreground region is obtained by the formulated foreground approximation detailed in the previous section. If the foreground approximation is accurate enough to localize the foreground object then the quality of the feature can be estimated by its pixels' intensity ratio inside and outside the foreground segmented area and the pixels' density inside the foreground segmented area.

The proposed feature density  $f_D$  is formulated as follows:

$$f_D = \underbrace{\frac{\sum_{\mathbf{p} \in \mathcal{P}} \mathbf{F}(\mathbf{p})}{\sum_{\mathbf{p} \in \mathcal{M}} \mathbf{F}(\mathbf{p})}}_{\text{Intensity ratio}} + \underbrace{\frac{\sum_{\mathbf{p} \in \mathcal{P}} \mathbf{F}(\mathbf{p})}{|\mathcal{P}|}}_{\text{Density term}}. \quad (8.3)$$

In (8.3),  $\mathcal{M}$  is a set of coordinates corresponding to all the pixels in (the feature map)  $\mathbf{F}$ . To be consistent with previous notation,  $\mathcal{P} \subset \mathcal{M}$  is defined as the set of coordinates of non-zero pixels in the segmented foreground approximation map  $\hat{\mathbf{F}}_G$ . The first term in  $f_D$  ensures that the intensity of  $\mathbf{F}$  inside the region defined by  $\hat{\mathbf{F}}_G$  is higher than the overall intensity of  $\mathbf{F}$ . The second term verifies that the portion of  $\mathbf{F}$  in agreement  $\hat{\mathbf{F}}_G$  has a higher feature density as compared with rest of the map.

The first row of Figure 8.3 shows an evaluation example of the proposed feature density cue. It can be noted that despite M1 assigning low intensity to the salient object as compared with the falsely highlighted background region, saliency index (SI) [49] and compactness (SV) [27] measures<sup>1</sup> rank it first due to its low spatial variance and high connectedness. On the other hand the proposed  $f_D$  cue ranks M2 first, which is in agreement with the F-max performance measure (for details of F-max, see section 8.4).

### Sparse Reconstruction Error Based Feature Quality

A good quality saliency map must be able to appropriately highlight all regions of the image that belong to the salient object. Predicting parts of

---

<sup>1</sup>These measures are defined in the background section.

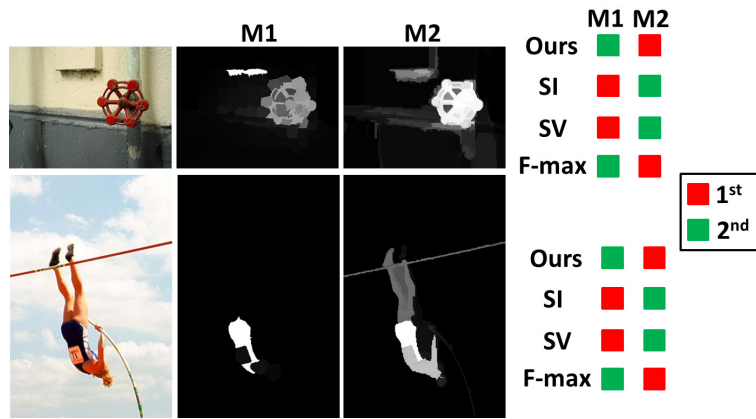


Figure 8.3: Rationale for feature density  $f_D$  and sparse reconstruction error  $\epsilon^s$  cues. From left to right: input image, example saliency maps and saliency map ranking. The red box depicts first rank for the respective map, while the green box depicts the second rank. M1 and M2 denote the saliency maps being assessed. Images are taken from the MSRA10K dataset (section 8.4).

the salient object as background results in undesirable object segmentation. This property of a saliency map is termed as salient object coverage, which is the number of total correctly predicted pixels that belong to the salient object. In this work, the extent to which a saliency map correctly covers all pixels of the salient object is measured by introducing a new feature quality measure based on the sparse reconstruction error between the saliency based predicted regions and the foreground approximation based predicted regions that were discussed in section 8.3.2.

The rationale behind the proposed measure is that it will discriminate between a saliency map that has a higher salient object coverage and another map having a lower coverage by assigning a lower total reconstruction error to the latter. Sparse reconstruction error has been used before by the Bayesian formulation of Li et al. [91] for constructing a saliency map, where the boundary nodes of an image were used for sparse representation. In this work it is employed to measure how well a saliency map highlights all the pixels belonging to the salient object.

To obtain the reconstruction error, each of the  $M$  regions is represented by mean colour features and pixel coordinates as  $\mathbf{x}_i = [R, G, B, L, a, b, x, y]^T$  and the image is described as a feature matrix  $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M] \in \mathbb{R}^{8 \times M}$ . Each feature based predicted segment is then extracted using its segmented version to construct a feature prediction set  $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N]$ . The foreground approximation  $\mathbf{F}_G$  based predicted segments  $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N]$  form the bases for the sparse representation. The reconstruction error for the  $i$ th segment  $\epsilon_i^s$  can then be computed as:

$$\epsilon_i^s = \|\mathbf{p}_i - \alpha_i \mathbf{S}\|_2^2, \quad (8.4)$$

where  $\alpha_i$  is the reconstruction coefficient given as:

$$\alpha_i = \arg \min_{\alpha_i} \|\mathbf{p}_i - \alpha_i \mathbf{S}\|_2^2 + \delta \|\alpha_i\|_1, \quad (8.5)$$

for  $\delta$  a regularization coefficient as in (8.5). The combined sparse reconstruction error  $\epsilon^s$  is then computed as  $\epsilon^s = \sum_i^M \epsilon_i^s$ .

Figure 8.3 shows an evaluation example of the total sparse reconstruction error  $\epsilon^s$  introduced in this work. The second row result shows a case, where the map M1 covers a small part of the salient object as compared with M2, which almost covers the whole salient object. Due to low spatial variance and high connectedness of M1, both SI and SV measures favour it by assigning it the first rank. The proposed total sparse reconstruction error  $\epsilon^s$  measure ranks M2 higher because of its higher coverage of the salient object as compared with M1. Consequently, the ranking of the proposed  $\epsilon^s$  measure is in accordance with the F-max measure.

### Quality of Feature Induced Segmentation

To induce a good segmentation result, a saliency map must be able to capture the high contrast between the “to be segmented region” and the rest of the map. To measure this property of a saliency map, a feature induced segmentation quality  $s_Q$  is introduced, which is inspired by the energy function of the ratio cut image segmentation method [135]. According to ratio cut, the segmentation should minimise the inter-region similarity, which is formulated in terms of the affinities at the segments boundaries as

$$s_Q = \frac{\sum_{p \in \mathcal{P}, q \in \mathcal{Q} \cap \mathcal{N}_p} w_{pq}}{\sum_{p \in \mathcal{P}, q \in \mathcal{N}_p} \mathbf{1}_Q(q)}, \quad (8.6)$$

where  $\mathcal{P} \subset \mathcal{M}$  is the set of coordinates of zero pixels and  $\mathbf{1}_Q(q)$  is the indicator function defined as

$$\mathbf{1}_Q(q) = \begin{cases} 1 & \text{if } q \in \mathcal{Q} \\ 0 & \text{if } q \notin \mathcal{Q}. \end{cases} \quad (8.7)$$

In (8.6),  $\mathcal{N}_p$  is the set of coordinates of neighbouring pixels to  $p$  and  $w_{pq}$  is the colour and spatial affinity between neighbouring pixels accessed by coordinates  $p$  and  $q$ .

To this end the affinity between neighbouring pixels is defined by [94]

$$w_{pq} = \exp(-\beta \|I_p - I_q\|)$$

$$\text{for } \beta = \left( 2\mathbb{E}(\|I_p - I_q\|^2) \right)^{-1} \quad (8.8)$$

The notation  $I_p$  in (8.8) represents the colour values of a pixel at the location  $p$  in the image  $I$ . Following from (8.6) and (8.7), equation (8.8) assumes that the affinities are only computed between neighbouring pixels such that the pixel at  $p$  belongs to  $\mathcal{P}$  and that at  $q$  belongs to  $\mathcal{Q}$ .  $\|\cdot\|$  represents the  $L_2$  norm. The  $\beta$  term in (8.8) computes the expectation over the selected pixel set (defined by (8.6) and (8.7)) in terms of colour contrast between neighbouring pixels.

In past work, Mai and Liu [99] used normalized cut as the energy function to measure the segmentation quality of a saliency map. In this work the energy function is designed to follow the ratio cut in contrast to normalized cut, due to its ability to judge a segmentation better than normalized cut as reported in [22] on the PASCAL VOC2009 training set. Furthermore, Wang and Mark demonstrate its superiority [135].

### Gestalt Laws of Contours for Feature Quality

As the introduced segmentation quality measure,  $s_Q$ , judges the boundary preserving capability of the segmentation result induced by the saliency map, a complementary measure is required that can measure the edge preserving quality of the saliency map. To this end the inter-region contour energy  $e_{\text{contour}}$  is computed according to Gestalt properties for contours with assistance from the proposed foreground approximation. The normalized sum of edge energies along the boundary of the contour is captured, which is a highly discriminative feature [22].

The inter-region contour energy  $e_{\text{contour}}$  can be expressed as follows:

$$e_{\text{contour}} = \frac{\sum_{p \in \mathcal{B}} E_I(p)}{|\mathcal{B}|}, \quad (8.9)$$

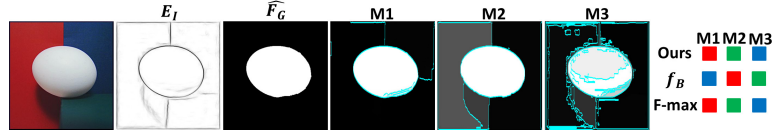


Figure 8.4: Comparison of the inter-region contour energy  $e_{\text{contour}}$  measure with  $f_B$  [99]. From left to right: image from the MSRA10k dataset (section 8.4), edge map  $E_I$ , segmented foreground approximation  $\hat{F}_G$ , selected saliency maps with overlaid boundaries and their respective rankings. Red, green and blue boxes represent first, second and third rank respectively.

where  $\mathcal{B} = \mathcal{B}_S \cap \mathcal{B}_F$  with  $\mathcal{B}_S$  and  $\mathcal{B}_F$  being the set of locations of boundary pixels of the saliency map and the foreground approximation, respectively and  $E_I$  is the edge map of image  $I$  computed using the approach of Dollár and Zitnick [36]. The approach of Dollár and Zitnick introduced random forests to learn structured class labels. Multiple forests are employed, where each predicts labels for a patch of edge pixels. The predicted labels are aggregated across the image to compute the final edge map.

The boundary quality measure of Mai and Liu [99] denoted as  $f_B$  compares the boundary map of the saliency map with its edge map. It can be misleading in scenarios where the background boundaries captured by a saliency map highly correlate with the image edge map. In contrast to Mai et al. [99], in this work the edge energy is only measured at that boundary of the saliency map that coincides with the boundaries predicted by the proposed foreground approximation map, thereby promoting edge energies at salient object boundaries instead of the background.

The discriminative nature of the proposed inter-region contour energy  $e_{\text{contour}}$  measure as compared with the boundary quality measure of [99] is depicted in Figure 8.4. It can be observed that the  $f_B$  measure favours M2 due to its high correlation with  $E_I$ . In contrast, the proposed  $e_{\text{contour}}$

measure promotes M1 (by taking assistance from  $\hat{F}_G$ ), which effectively captures the edges of the salient object and suppresses background edges. The F-max measure validates the proposed  $e_{\text{contour}}$  measure.

### Systematic Evaluation of the Proposed Quality Measurement Cues

To investigate the ability of the proposed quality measurement cues to discriminate between good and bad saliency maps, 8000 saliency maps were computed using images from the MSRA10K dataset (section 8.4) and six state-of-the-art saliency methods namely, AC [2], CA [46], FT [3], GBVS [50], HC [28] and RC [28]. Next, the maps were divided into high performing (high) and low performing (low) categories based on their F-max measure. The proposed measurement cues were computed for both the high and low performing maps and their respective distributions were plotted, as depicted in Figure 8.5. Two-sided Wilcoxon rank sum tests were performed to compare the high and low distributions. The distributions were found to be highly statistically different with p-values  $< 0.0001$  for all the measurement cues.

#### 8.3.4 Objective Function for Feature Quality

To get a collective assessment of quality of a map, information from all the quality measurement cues must be combined into a single objective function. As all feature quality cues are unnormalized, they are first normalized to the range  $[0,1]$ . Afterwards, the objective function for computing the quality of the  $i$ th feature is formulated as:

$$t_i = \frac{f_D + e_{\text{contour}} + s_Q}{\epsilon^s}. \quad (8.10)$$

The idea is that all the measures that must assume high values to indicate a good salient object detection result are placed in the numerator and the sparse reconstruction error that should be low to indicate a good



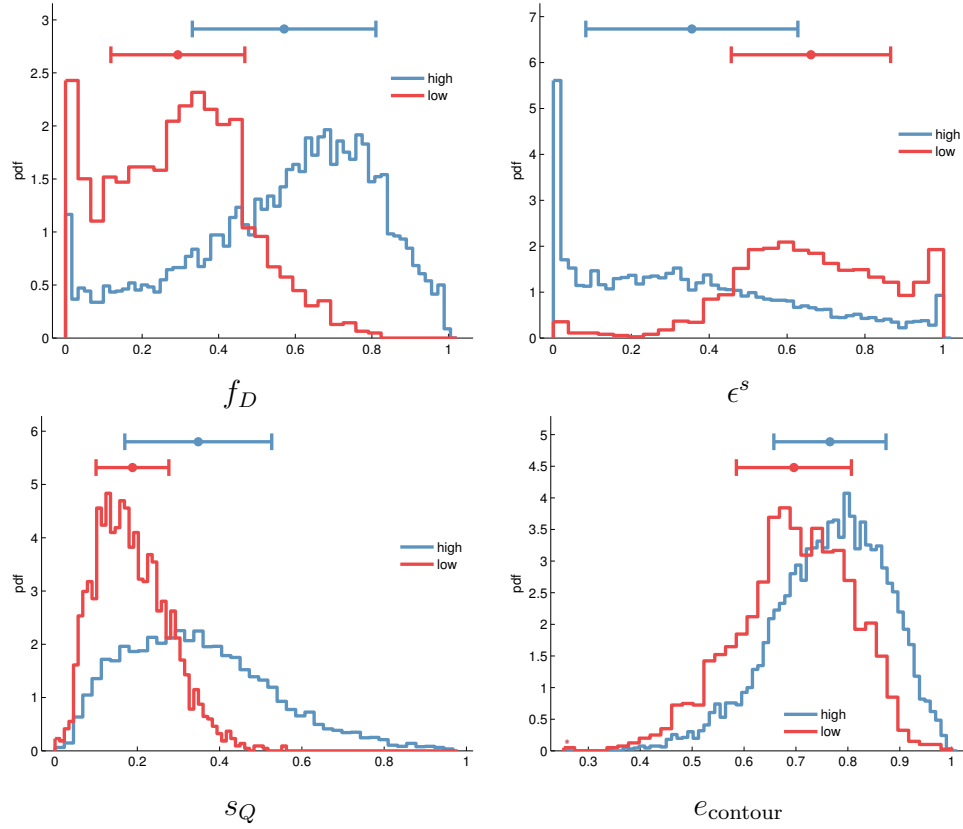


Figure 8.5: Distributions of the proposed quality measurement cues for high and low performing maps. The high and low performing maps are segregated based on their F-max measure.

salient object result is placed in the denominator. Other more sophisticated objective functions including weighted combination were investigated. However, it was observed that the seemingly simple quality measure as in (8.10) performs effectively as demonstrated by the results in sections 8.4, 8.6 and specifically the ranking evaluation 8.7. Also, the potential gains from more complex approaches were questionable.

### 8.3.5 Feature Selection Algorithm

The proposed DFS method can be utilised in two different approaches of operation for feature selection. While operating in the first approach, the proposed DFS method selects the subset of best performing features on an image-by-image basis as described by Algorithm 3. The feature maps set  $\mathcal{F}$  and quality vector  $\mathbf{t}$  are computed for each image and passed as input to the algorithm.  $n_c$  is the number of clusters to be formed by hierarchical clustering in our implementation.  $\mathcal{F}_s$  is the set of selected features containing feature maps  $\mathbf{F}_s$ , while  $\mathbf{DFS}$  is the saliency output for the proposed DFS method.

The best features for the fusion process are selected by hierarchical clustering of features based on their respective qualities given by the feature quality vector  $\mathbf{t}$ , which contains sorted feature qualities predicted by our feature quality measurement cues. The idea is to cluster the high quality features (having close proximity in terms of feature qualities) to form the selected feature set  $\mathcal{F}_s$ . Pairwise Euclidean distance is used to obtain the agglomerative hierarchical cluster tree. The cluster tree is then used to obtain two clusters based on a the specified number of clusters  $n_c$ . The cluster containing the highest accuracy is then selected to populate the selected features set  $\mathcal{F}_s$ .

The process of grouping good and bad features using agglomerative hierarchical clustering for a representative case is depicted in Figure 8.6. The process of hierarchical clustering is simple yet effective. It starts by as-

---

**Algorithm 3:** Dynamic feature selection on an image-by-image basis.

---

**Data:**  $\mathcal{F} = [f_1, f_2, \dots, f_N]$  - feature/saliency method set  
corresponding to an image,  $t = [t_1, t_2, \dots, t_N]$  - feature quality  
vector

**Result:**  $DFS$  - DFS Saliency

Sort features  $f_i \in \mathcal{F}$  according to  $t$  in descending order ;

Compute distance matrix  $D$  for  $t$  using linkage ;

**while**  $n_c > 1$  **do**

    Merge clusters having closest distance to form the hierarchical  
    clustering tree;

    Use inter-cluster distances to reduce the dimension of  $D$  ;

Create clusters by specifying the number of clusters required ;

Use the cluster label with higher accuracy to populate the selected  
feature set  $\mathcal{F}_s$  for the current image  $I$  ;

Optimally combine selected feature maps  $F_s \in \mathcal{F}_s$  to obtain  $DFS$ ;  
**return**  $DFS$  ;

---

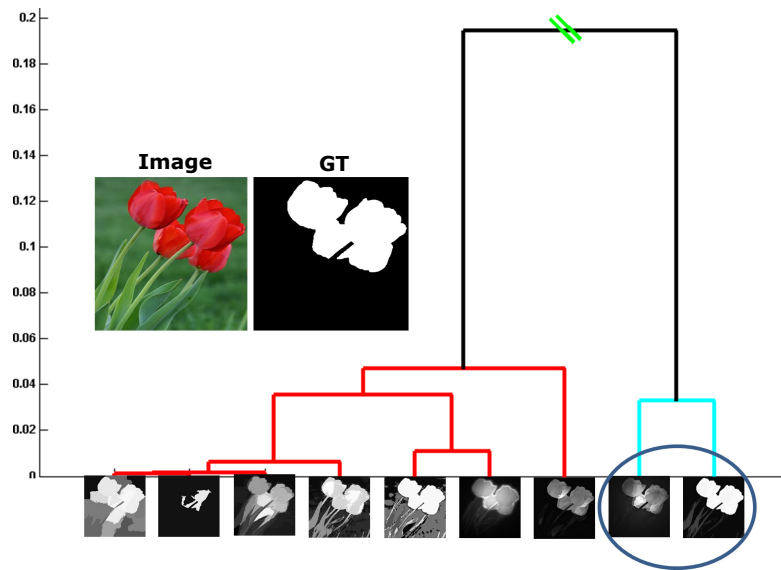


Figure 8.6: Example of an agglomerative hierarchical cluster tree for selecting appropriate features for combination. The red, blue and black colours are assigned to group of nodes whose distance is less than a threshold. The colour assignment is unique and random. The green lines depict the points where the dendrogram is split into two distinct clusters.

signing each object to a separate cluster followed by computation of pairwise distances between clusters. Euclidean distance is found to be highly effective in our implementation and is therefore employed to evaluate distances. Using all distances, a distance matrix is constructed and pair of clusters with shortest distance are sought. Next, such pairs are merged and removed from the distance matrix. Matrix is updated by computing distances of all clusters from this newly created cluster. This process is repeated until the distance matrix is reduced to a single element. In Figure 8.6, the height of the “U” shapes depicts the euclidean distance between objects, while the green lines signify joints from where the hierarchical tree is cut to separate good and bad features. The blue ellipses indicate features that are selected by the dynamic feature selection method for combination based on their quality. More examples of clustering features for multiple combination schemes are presented in chapter 9.

In the second approach, the proposed DFS approach can be used by saliency methods in an offline mode, and by large-scale feature integration systems to identify such features that add to the computational load (of the system) without improving performance.

### 8.3.6 Feature/Saliency Fusion

As the prime motivation for this work is feature selection, basic fusion schemes from the literature are employed in this work instead of using more sophisticated multiple learners approaches such as [100, 104]. For details of these approaches, please refer to the background chapter.

To fuse the saliency maps, we adapt the pixel-wise saliency aggregation approach of Mai et al. [100]. Specifically a feature vector  $\mathbf{f}_p$  is formed as  $[s_1, s_2, \dots, s_m]^T$  (considering  $m$  saliency methods;  $s_i$  is the saliency of a pixel according to the  $i$ th method) with a corresponding binary label vector  $\mathbf{l}_p$  to encode saliency of pixels. The final saliency  $s_p$  for a pixel is determined by finding the posterior probability of the pixel being salient

or not. Specifically, the posterior probability is modelled using the logistic method as follows:

$$P(\mathbf{l}_p = 1 | f_p; \boldsymbol{\lambda}) = \sigma \left( \sum_{i=1..m} \lambda_i s_i + \lambda_{m+1} \right), \quad (8.11)$$

where  $\sigma(\cdot)$  is the sigmoid function and  $\boldsymbol{\lambda}$  is a vector containing model parameters [Note that this is different from the balancing parameter  $\lambda$  introduced in section 8.3.2]. For learning model parameters  $\boldsymbol{\lambda}$ , a non-linear SVM with RBF-kernel is implemented using the publically available LIB-SVM package [24]. These settings employed in this work are likely to be different to the settings used in [100] as the authors do not report the parameters used.

In order to combine the multi-level saliency maps in the benchmark DRFI method, a least square estimator is employed as reported in [68] to learn the weights for saliency maps by minimizing the sum of the losses over the training images.

## 8.4 Experimental Setup

### Data Sets

The **MSRA10k** database [28] contains 10,000 real-world images and is a subset of the MSRA salient object database [94]. Its prime feature is the pixel-level ground truth masks, which allows finer grained evaluations than the simpler bounding box annotations provided by MSRA. Due to its large size and variety of image types, it can test the scalability of saliency methods.

The **ASD** dataset [3] contains 1000 images with pixel-level ground truth annotations. It is a subset of the MSRA10k dataset where most images contain a single object with a simple background, which makes it a relatively easier dataset.

The **BSD300** dataset [102] also known as SOD in prior works [15] is a subset of the Berkeley segmentation dataset, containing 300 images in total, labelled by seven users as object boundaries. It will be referred as BSD for the remainder of this chapter. This dataset is relatively more difficult than MSRA10k and ASD as it contains images with unobtrusive objects and cluttered background.

The **SED2** dataset is a subset of the segmentation evaluation database [9]. It comprises 100 images containing two objects and was originally collected to evaluate image segmentation algorithms. Pixel-level ground truth annotations from at most three users are also provided. The inclusion of multiple objects per scene makes it challenging for most saliency methods.

The **PASCAL VOC 2012** consists of images obtained from the Flickr website [39] for the PASCAL VOC challenge. The images are collected over the years from 2007 for the challenging tasks of classification, detection, segmentation, action classification and persons layout, including 20 object categories. This work employs 2,913 images out of the total 11,540 images, for which the segmentation annotations are made available. Complete annotations are provided for the twenty classes with bounding boxes and attributes specifying the objects, actions and layouts. Pixel-wise ground truth is available for the segmentation task, where each pixel is annotated according to its object class level. Binary ground truth is obtained for pixel-wise segmentation by assigning “one” to each pixel belonging to any of the twenty object categories and assigning the rest of the pixels “zeros”. PASCAL VOC is one of the most difficult benchmark dataset and is used as a measure of assessing state-of-the-art performance in a competition since more than five years.

As this work presents the first feature selection method introduced in this thesis, the MSRA10k, BSD and the SED2 datasets are employed to thoroughly evaluate the performance of methods. In order to be consistent with the results reported in [100], the ASD dataset is employed instead of

the MSRA10k dataset for the saliency aggregation experiments and comparisons. For accessing the performance of DFS based salient object detection method proposed in this work, the PASCAL VOC 2012 dataset is employed (in section 8.6) to maintain consistency with prior works.

## Evaluation Metrics

All methods are evaluated using the average precision-recall (PR) curves. Using the FT benchmark [3], each saliency map is first subjected to segmentation using thresholds in the range  $[0,255]$  and compared to the ground truth annotation to compute the precision and recall at each threshold level. The average precision and recall for all images are then used to plot the PR curve. In addition to the PR curve, a segmentation inspired measure termed as F-max is also included for comparison. It is included as indirect measure of the quality of segmentation induced by the selected features. It is computed as the maximum F-measure value when operating on the PR curve. F-measure is computed by the same Equation 5.8 as before, where  $\beta$  is set to 0.3 to reflect the FT benchmark [3]. For limited space in figure labels, F-max is denoted by the letter F only. The contours for F are also plotted in the figures for better illustration.

## Parameter Analysis

This sections presents an empirical analysis of various parameters of the proposed method. There are three parameters in the foreground approximation ( $F_G$ ) computation, number of regions  $M$ , strength of the edge weight  $\sigma$  (6) and the balancing parameter  $\lambda$  (8.2). Figure 8.7 shows the effects of these parameters on the quality and the computational speed of  $F_G$ . These results are computed by using 100 randomly sampled images from MSRA10k dataset. In Figure 8.7 (a) the effect of  $M$  on the perfor-



mance of  $F_G$  is plotted in terms of F-max, mean absolute error<sup>2</sup> (MAE) [113] and computational time in seconds. It can be observed from Figure 8.7 (a) that with the increase in  $M$ , there is a gradual increase in computational time, while F-max and MAE do not exhibit considerable variation. F-max and MAE show an increasing and decreasing trend till  $M = 200$ , while computational time increases steadily. In contrast to the low variations in F-max and MAE on increasing  $M$  above 200, there is a noteworthy increase in the computation time having a steeper slope than the curve before  $M=200$ . Therefore  $M$  is set to 200 to favour a balance between quality and computational speed.

It can be observed that varying  $\sigma^2$  and  $\lambda$  do not have a profound effect on the computational time, while F-max and MAE both exhibit best scores for  $\sigma^2$  fixed to 0.05 (corresponding to all  $\lambda$  values, see Figure 8.7 (b)-(d)). Hence  $\sigma^2$  is fixed to 0.05, while  $\lambda$  is set to 0.99, as they result in the best F-max and MAE.

In our experiments,  $\delta = 0.03$  is found to be a good value for estimating the reconstruction coefficient in (8.5). After performing a parametric search on the grid the values of  $C$  and  $\gamma$  for the non-linear SVM, they were fixed to two and eight respectively, due to their high cross-validation accuracy. The number of clusters  $n_c$  is set to two in our implementation as ideally two clusters containing good and bad performing features are required.

## 8.5 Results and Discussion

This section presents the comparison of the results obtained by the unique feature combinations obtained by the proposed DFS method as compared with the optimal combination of all features using three different exper-

---

<sup>2</sup>MAE is an important performance measure that measures the correctness of background annotations in saliency maps, which is usually neglected by other benchmarks [113].

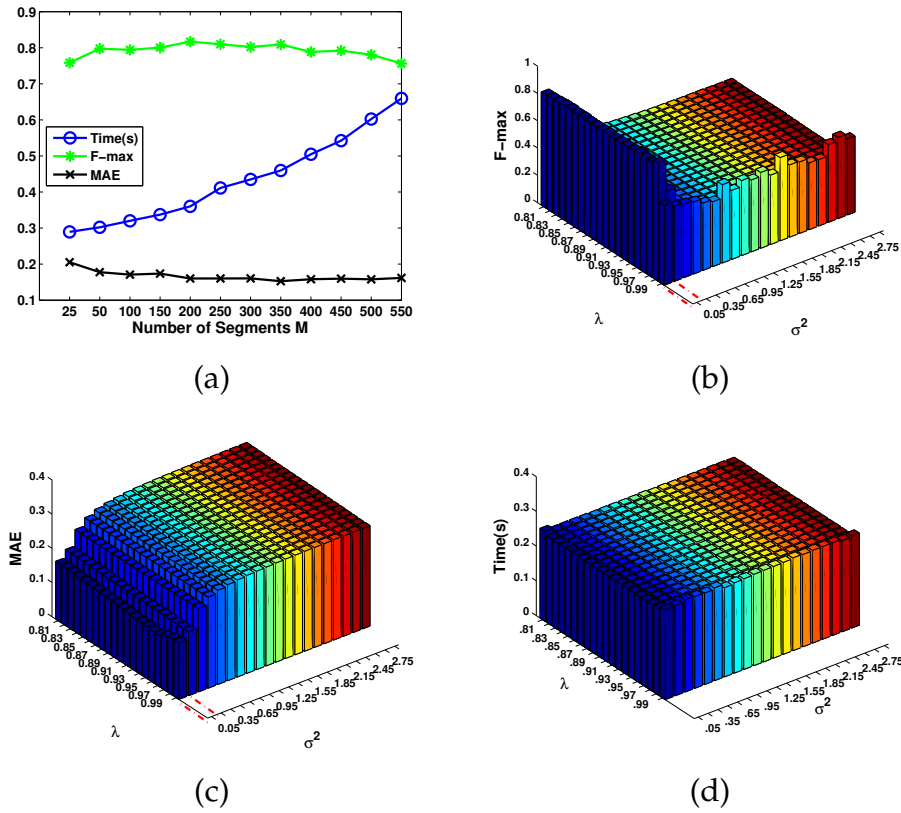


Figure 8.7: Effects of parameters on the foreground approximation  $F$ . (a) Quality and speed of  $F$  with varying number of regions. (b)-(d) Effects of varying  $\sigma^2$  and  $\lambda$  on F-max, MAE and time in seconds, respectively.

iments. Firstly, the performance of the proposed DFS method is evaluated by employing it to autonomously select and combine the multi-level saliency maps of the DRFI method. The performance of the proposed DRFI method is compared with the multi-level saliency fusion of the DRFI method, which combines all the saliency maps. Next, the saliency aggregation performance of the proposed DFS method is compared with the state-of-the-art saliency aggregation method, i.e. PW [100] to investigate the benefit of unique saliency combinations as compared with the optimal combination of all saliency maps. Finally, the performance of the proposed DFS based salient object detection method (section 8.6) is compared with the state-of-the-art salient object detection methods. A comparison of the proposed DFS based method with the best selected map for each image according to the DFS quality is also presented.

### 8.5.1 Saliency Map Selection for Regional Feature Integration

The performance of the proposed DFS method is computed by employing it at the multi-level saliency stage of the DRFI method [68] to perform autonomous saliency selection and fusion. For the DRFI method, its final output after the multi-level saliency fusion stage is used. The MATLAB implementation provided by the authors of DRFI [68] is employed in this work.

The proposed DFS method automatically selected less up to five features for each image of the MSRA10k and SED2 datasets and up to four features for each image of the BSD dataset. Figure 8.8 shows the performance of DRFI and the proposed DFS method along with the performance for dynamically selected best and worst performing saliency maps. In comparison with the DRFI method the proposed method shows similar performance on the MSRA10k dataset, slightly better performance on the SED2 dataset and notably better results on the BSD dataset in terms

of the PR curves. The results of the PR benchmark demonstrate that the proposed DFS method improved the performance of the benchmark DRFI method by discarding poor performing saliency maps at its multi-level saliency fusion stage. Notably, the dynamically selected best map shows better performance than baseline DRFI on two datasets in terms of PR curves. In order to further quantify the results, the average area under the ROC curve (AUC) with the statistical significance results of a paired, two sided, Wilcoxon signed rank test are included. The results show that the proposed DFS method performs significantly better than DRFI on the BSD dataset with a p-value  $< 0.0001$ , while showing statistically similar performance on the MSRA10k and the SED2 datasets with p-values of 0.7985 and 0.9288, respectively.

Figure 8.9 shows a visual comparison of the proposed DFS method with the DRFI method. Figure 8.9 includes two representative saliency maps from the set of 15 multi-level saliency maps of DRFI. The second and third row images from the MSRA10k and SED2 datasets show that the proposed DFS method and the baseline DRFI method perform similarly (supporting the PR curves). The image in the first row from the BSD dataset shows improvement of the proposed method over the baseline DRFI method in that the DFS method uniformly highlights the salient region (the flowers) including some background portion of the building, while the DRFI method completely misses the salient region in the final saliency output.

As the number of features selected on all datasets is less than or equal to five, the results of the proposed DFS method suggest a potential speed-up of approximately three times over DRFI. The overall results of the comparison suggests that the proposed DFS method either improves the performance of DRFI or provides similar quantitative performance (on a few datasets) with the potential of low computational cost that can be achieved by discarding the outlier saliency maps.

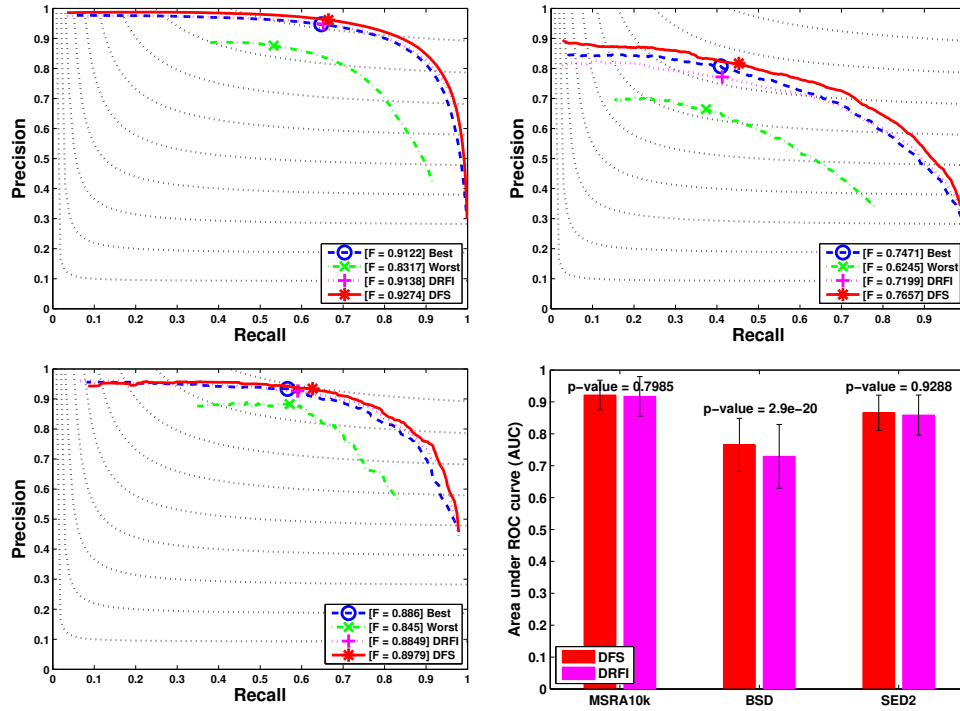


Figure 8.8: Top row: precision-recall curves for multi-level saliency fusion on the MSRA10k and BSD datasets; bottom row: precision-recall curves on the SED2 dataset and statistical comparison based on AUC. Best and Worst represents the dynamically selected best and worst maps for each image from the complete set. The proposed DFS method is shown in Red. This figure is best viewed in colour.

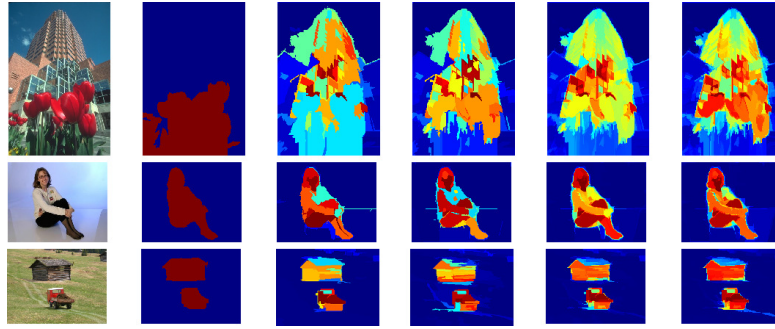


Figure 8.9: Saliency comparison results of two representative multi-level saliency maps of DRFI (M1 and M2), DRFI and the proposed DFS method on representative images taken from BSD, MSRA10k and the SED2 datasets. Left to right: input, GT, M1, M2, DRFI and DFS.

### 8.5.2 Saliency Aggregation

The ten state-of-the-art saliency methods AC [2], CA [46], IT [66], MZ [98], LC [150], FT [3], HC [28], GBVS [50], RC [28] and SR [55] are employed, as in [100]. Specifically the FT benchmark [3] is utilised to generate saliency maps for all the methods reported here. The proposed DFS method is compared with the pixel-wise saliency aggregation method of Mai et al. [100] (termed as PW) to investigate the potential of selective feature fusion in comparison with the combination of all features.

The proposed DFS technique selected less than five saliency methods from the 10 methods for every image belonging to all three datasets. Figure 8.10 shows the quantitative performance of methods, where the proposed DFS method performs better than PW on all the three datasets, achieving higher precision and recall on all thresholds as supported by the statistical results on the AUC. The statistical results were obtained by a paired, two sided, Wilcoxon signed ranked test.

Figure 8.11 presents the visual comparison for the task of saliency aggregation. It can be noted that the proposed DFS method predicts better

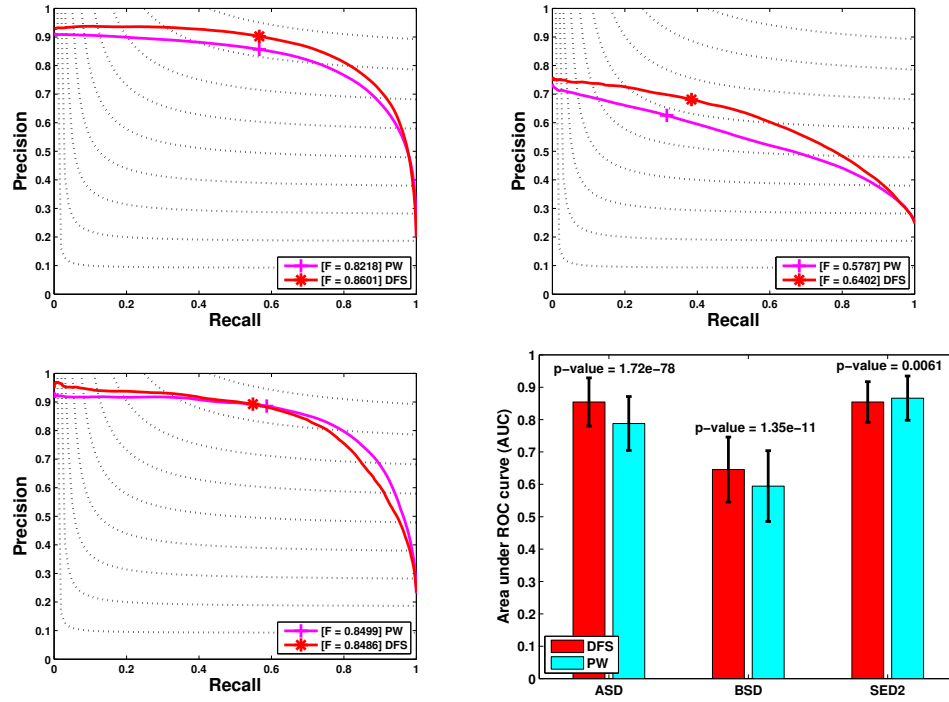


Figure 8.10: Top row: precision-recall curves for saliency aggregation on the ASD and the BSD datasets; bottom row: precision-recall curves on the SED2 dataset and statistical comparison based on AUC. The proposed DFS method is shown in red. This figure is best viewed in colour.

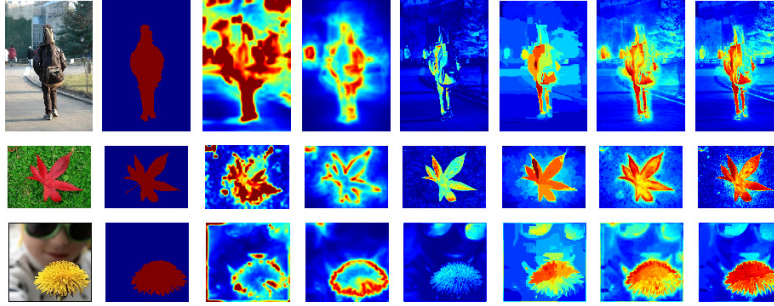


Figure 8.11: Saliency comparison of the proposed DFS with the PW method and selected individual saliency maps. Left to right: input, GT, GBVS, CA, FT, RC, PW and the proposed DFS.

saliency as compared with the PW method and all individual methods.

The results of the multi-level saliency fusion of the DRFI method (section 8.5.1) and the results of the saliency aggregation of methods (section 8.5.2) are in agreement with our initial hypothesis. This is because the fusion of dynamically selected features by the proposed DFS method improves upon the methods that have to use all the features. The obvious reason for this improvement is the rejection of redundant saliency maps by the DFS method.

## 8.6 Results of the DFS Based Saliency Aggregation Method for Salient Object Detection

This section introduces a DFS based saliency aggregation method for the task of salient object detection. It employs five saliency methods namely, FT [3], HC [28], RC [28], SR [55] and DRFI [68] that have been discussed in section 8.5.2. Furthermore, two additional methods are included in the design, i.e., GBMR [149] and PCA [101], due to their good performance on benchmark datasets. Author provided implementations are used to obtain



the saliency maps. The proposed DFS method is applied to dynamically select and combine the best possible saliency maps on individual image basis from the PASCAL VOC 2012 dataset.

Additionally, the sub-class(es) of images for which the proposed DFS based saliency method is best suited is investigated. Figure 8.12 shows the performance comparison of the proposed DFS saliency method with the other saliency methods. Along with the performance of individual saliency methods and the proposed DFS based saliency approach, the average performance of only the best saliency method (selected by DFS) per image, termed as Best is also plotted. It can be observed from the precision-recall curves (in the left column) that the proposed DFS based saliency method achieves considerably better precision and recall as compared with the benchmark *learning based feature aggregation method*, i.e., DRFI on all thresholds. The performance enhancement of the proposed method over the state-of-the-art methods is pronounced as compared with the lowest performing saliency method FT. Additionally it can be seen that the proposed method achieves considerably better performance as compared with Best, as it has the ability to select the set of best performing maps depending upon image type and combine them optimally to achieve the desired results.

### 8.6.1 Further Discussion

The precision-recall curve in the middle column of Figure 8.12 shows the performance of methods on a subset of images, where the proposed DFS based saliency method performs better than all the saliency methods. It can be observed that the high performing methods, i.e, Best, DRFI, GBMR, PCA and RC, and the low performing methods namely, FT, HC and SR are tightly clustered on this subset of images in terms of their performance with a considerable gap between the high and low performing methods. It is hypothesized that the images that belong to this class are difficult

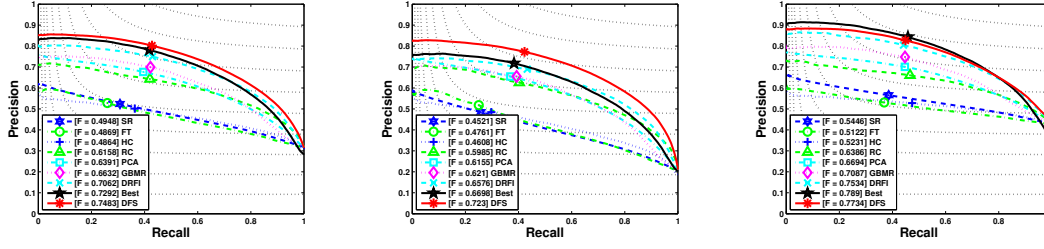


Figure 8.12: Left: precision-recall graph for saliency aggregation on the PASCAL VOC dataset. Middle: precision-recall curve for images on which the proposed DFS based method outperforms all other saliency methods. Right: precision-recall curve for images on which the proposed method is not the best among methods.

images, hence there is a considerable difference between the performance of low and high performing methods.

The precision-recall curve on the right of Figure 8.12 shows the performance of methods on the subset of images where the proposed method was not the best performer among all the methods. The uniform spread in the performance of methods on this subset of images makes it difficult for the DFS based saliency approach to differentiate between maps and select the best set for combination.

Upon visual analysis of the images where the proposed approach performed better than the benchmark methods, most of these images could be categorised as either having cluttered backgrounds or multiple salient objects. This is in agreement with our initial hypothesis. Representative examples of images from both classes are shown in Figure 8.13. The first three rows show images with cluttered backgrounds and the last three rows provide example images with multiple salient objects. It can be observed from the first and third row images that the proposed DFS based saliency method selects and combines the best maps to appropriately highlight the salient objects and capture their shape. For multiple objects, the proposed method produced saliency maps covering more objects with



Figure 8.13: Representative images are presented, where the proposed approach performed better than the benchmark methods. From left to right: input, GT, DRFI, FT, GBMR, HC, PCA, RC, SR, the proposed DFS based saliency and the precision-recall curve for each image. All images are taken from the PASCAL VOC dataset.

uniform highlighting as compared with the individual saliency methods. The pronounced response of the proposed DFS based saliency method on multiple objects can be observed from the image in the last row.

## 8.7 Ranking Evaluation

To fulfil the final objective of this work, this section evaluates the performance of the proposed DFS method in terms of its feature/saliency ranking capability as compared with the ground truth ranking and other standard ranking measures. Two metrics have been employed in this work to access the ranking results namely, Spearman's rank correlation ( $\rho$ ) [124] and Kendall's tau ( $\tau$ ) [73].

Spearman's rank correlation assesses individual rankings by finding the degree of a monotonic relationship between the variables. To compute Spearman's rank, individual ranks are first normalized by replacing the

rank for identical scores by the mean of ranks. The rank correlation is then computed as:

$$\rho = 1 - \frac{6 \sum (x_i - y_i)^2}{n(n^2 - 1)}, \quad (8.12)$$

where  $x_i$  and  $y_i$  are the normalized ranks and  $n$  is the size of the sample.  $\rho$  measures the correlation in rank order. A positive or negative value of  $\rho$  corresponds to an increasing or decreasing monotonic trend between the variables, respectively.

The Kendall's  $\tau$  measures the association between two random variables by penalizing discordant ranking pairs and promoting concordant ranking pairs. Given two ranking functions  $r_a$  and  $r_b$  and sets of observations  $X = \{x_i\}_{i=1..n}$  and  $Y = \{y_i\}_{i=1..n}$ , the Kendall's  $\tau$  rank correlation is given as follows:

$$\tau = 1 - \frac{2 \sum_{(i,j)} \delta \left( \text{sgn} \left( r_a(x_j) - r_a(x_i) \right), \text{sgn} \left( r_b(y_j) - r_b(y_i) \right) \right)}{n(n-1)}, \quad (8.13)$$

where  $\delta(\cdot, \cdot)$  is the Kronecker delta function. The two sign functions and the Kronecker delta ensure that all concordant ranking pairs of  $X = \{x_i\}_{i=1..n}$  and  $Y = \{y_i\}_{i=1..n}$  are accumulated resulting in an increase of  $\tau$ .

As the methods are assessed by a precision-recall curve on the benchmarks for salient object detection, the area under the precision recall curve (AUCPR) score is the most suitable ranking measure. Therefore, it is employed as the ground truth (ideal) ranking measure. To compare the proposed DFS based ranking results with other quantitative measures that were previously employed by saliency methods to judge their results, the area under the ROC (AUC) curve and the classification accuracy (CACC) measures are used. The AUC measure is frequently used as a standard measure to evaluate salient object detection results. For each saliency/feature map, the AUCPR, AUC and CACC measures are computed as per the FT benchmark [3]. The saliency/feature map is thresholded using multiple thresholds and compared with the binary ground truth to compute the

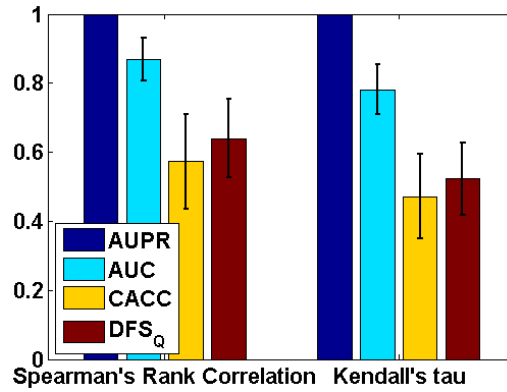


Figure 8.14: Ranking performance comparison of the proposed DFS method with the ideal ranking (AUCPR), AUC and CACC in terms of Spearman's Rank Correlation ( $\rho$ ) and Kendall's tau ( $\tau$ ).

precision-recall curve for AUCPR, the ROC curve for AUC and the classification accuracy curve for CACC. The area under the curves gives the corresponding metrics. For the proposed DFS method, a feature/saliency map is evaluated by the objective function for feature quality as in (8.10), termed as DFS<sub>Q</sub>.

From the ranking results in Figure 8.14, it can be seen that the proposed DFS<sub>Q</sub> is slightly better than CACC, with AUC performing the best

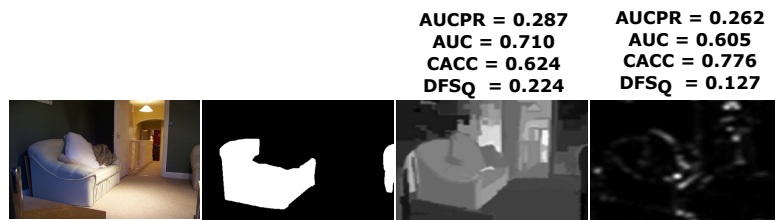


Figure 8.15: Representative visual example of ranking results. From left to right: input image from the PASCAL VOC dataset, GT, RC and SR overlaid with performance measures. DFS<sub>Q</sub> is normalized in the range [0,1].

with lowest deviation from the mean. As CACC encodes the number of correctly predicted pixels as compared with the total number of pixels, it can be misleading for the task of salient object detection. Figure 8.15 shows that CACC prefers the SR map (as the number of correctly predicted examples for SR are higher as compared with RC), ranking it higher than RC, opposing the AUCPR and AUC measures. It is to be noted that the proposed  $\text{DFS}_Q$  correctly ranks RC higher than SR as the aim of the quality  $q$  in the proposed DFS method is to assist salient object segmentation.

## 8.8 Chapter Summary

The aim of this chapter was to compare the salient object detection performance of a method that can autonomously select and combine the best performing maps with another method that combines all the maps for computing the final saliency output.

To achieve this goal, a method for dynamic feature selection (DFS) was introduced, which was able to measure the quality of a saliency map based on novel cues. A clustering technique was devised to cluster the best performing and unwanted maps. The performance of the proposed DFS method that employed unique feature combinations was compared against the performance of methods that combine all the maps. The results of the comparison suggested that the proposed method was able to separate good and bad performing maps during combination and therefore improved upon the salient object detection performance of methods that combined all the maps. Additionally, the proposed DFS based salient object detection method exhibited better performance than both learning and non-learning based state-of-the-art methods and also performed better than the single best map for individual images. The important findings of this work are as follows:

The results of the multi-level saliency fusion and the saliency method aggregation experiments suggested that the proposed method was able to

identify poor performing maps, which were falsely included in the fusion schemes of DRFI and PW, thereby improving upon their salient object detection performance.

The results of the DFS based salient object detection method revealed that the best map selected by the proposed DFS quality measure performed better than individual state-of-the-art methods supporting the feature selection paradigm adopted in this work. Additionally, analysing the subset of images on which the proposed DFS method performs better than all state-of-the-art methods suggested that it is due to the clustering of method performances on difficult images that helps the proposed method achieve effective feature selection and obtain improved performance. Also the visual analysis of that subset revealed that the difficult images on which the proposed methods outperform the state-of-the-art can be categorised as scenes with cluttered backgrounds and/or multiple salient objects.





## Chapter 9

### Discussions

A variety of methods to improve generalisation performance on unseen difficult images for salient object detection have been developed (chapters 3-8). These techniques include joint optimisation within a single combination scheme, multiple feature combination schemes and the use of dynamic feature selection before feature combination. The relationship amongst the various proposed methods, how they improve the current state-of-the-art and recommendations for incorporating them into current and future salient object detection methods are provided in this chapter. The overall structure of this work is presented in Figure 9.1.

The initial methods sought to combine all feature maps according to the learned combination schemes. These methods effectively control the contributions of features through assigned feature importance weights. Specifically the multiple combination schemes are highly general in defining the contribution of features to the final saliency. However, in the case of difficult images, undesired features exist that are better ignored during feature combination. Hence a method is needed that can aid the learning based combination methods by neglecting unwanted maps that may corrupt the final output. The dynamic feature selection method (DFS) (introduced in chapter 8) has the capability to select the best possible maps during combination and discard the noisy ones. The DFS method is adopted

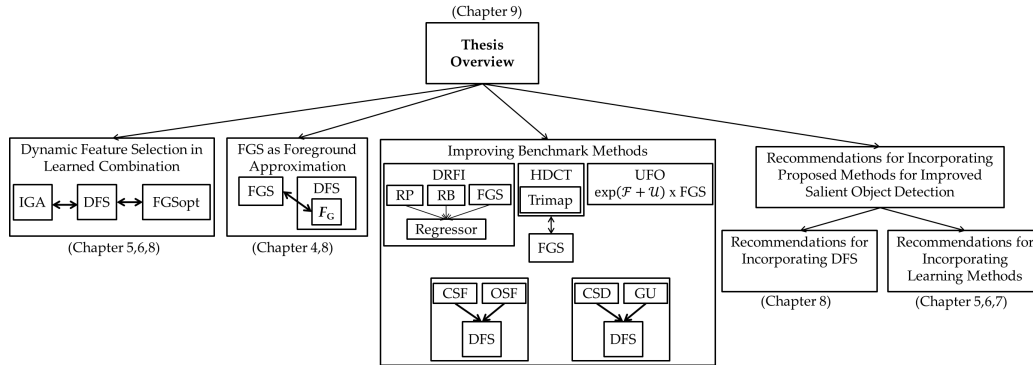


Figure 9.1: Overview of the thesis structure. The interconnections between the various methods introduced in this thesis are presented. For details about new acronyms and symbols, please refer to the text in individual sections.

to aid the learning based combination schemes, namely FGSopt, IGA and XCSCA (introduced in chapters 5-7).

A region based foreground approximation was introduced in the DFS method to complement the region based design of most benchmark saliency methods (see chapter 8). In contrast, a matting based FGS method to improve upon region based salient object detection approaches was introduced in chapter 4. FGS was investigated as an alternative foreground approximation for DFS in this chapter and trade-off between performance improvements and computational time is discussed.

The FGS method was introduced to overcome the limitations of region based benchmark methods (see chapter 4). The results suggested that the FGS method was successful in overcoming the problems of non-uniform saliency assignment and falsely highlighted background. This chapter investigates that whether the smooth and uniform response of the FGS method can aid the region based benchmark methods in overcoming their inherent artifacts. The DFS method was demonstrated to benefit saliency aggregation in chapter 8. However, whether it can benefit complementary feature selection was not investigated. This chapter explores that how the

DFS method can improve complementary feature selection in benchmark methods.

Finally, a set of guidelines are presented and implementation aspects of incorporating the proposed techniques in current and future salient object detection methods are discussed. Depending upon the application and requirements, it is discussed that how the variants of the proposed methods can be incorporated into other salient object detection methods.

## 9.1 Dynamic Feature Selection in Learned Feature Combination

The learned combination based methods including FGSopt, IGA and XCSCA seek to combine all available feature maps after considering their effect on the final saliency output. Taking into account all three approaches, this is achieved by proper feature conditioning, appropriate feature importance assignment and suitable integration of features. Specifically, the multiple rule based methods generalise well on a wide variety of image types to ensure appropriate combination with effective suppression or conditioning of noisy features. However, for difficult images, it becomes cumbersome for even the multiple combination rule based methods to properly condition and control the contribution of unwanted noisy features. In such cases, the dynamic feature selection method proposed in this thesis has the ability to aid the learned combination methods as discussed in chapter 8. However, no investigation has previously been performed on how the DFS method can aid the learned feature combination based methods to ignore unwanted maps during combination. This section investigates the role of the DFS method when incorporated in the learning based feature combination methods. Specifically, the DFS method is investigated in conjunction with the joint optimisation technique, i.e. FGSopt (introduced in chapter 5) and the IGA method, which is chosen as

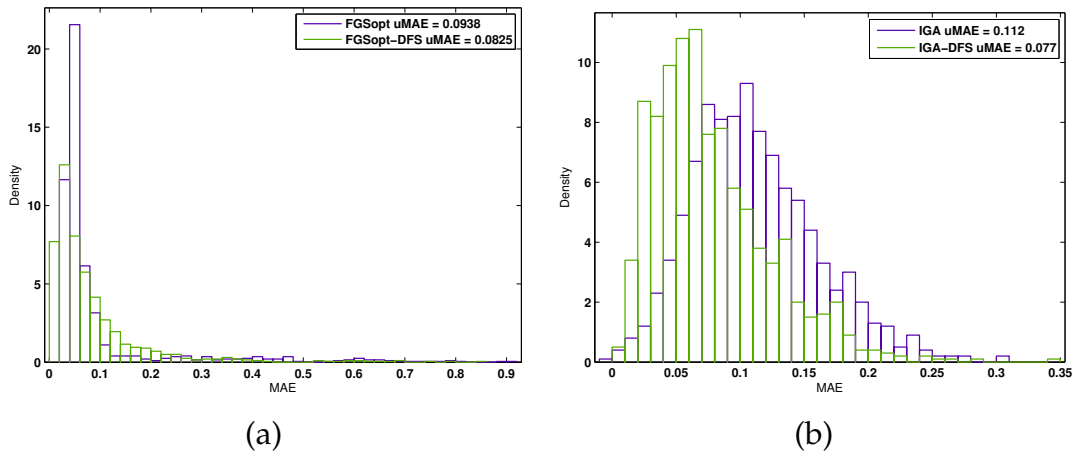


Figure 9.2: (a) Comparison of FGSopt and FGSopt-DFS in terms of average MAE on 1000 images of the ASD dataset. (b) Comparison of IGA and IGA-DFS methods in terms of average MAE on the ASD dataset.

a representative of multiple combination rules based methods. The modified FGSopt method with DFS incorporated is termed as FGSopt-DFS, while the modified IGA method with added dynamic feature selection capability is referred as IGA-DFS. It is anticipated that due to the ability of the DFS method to exclude noisy features during combination, the background noise suppression ability of the combination techniques will be improved. With this in mind, mean absolute error (MAE) is employed to evaluate how well the models predict the background regions in comparison with the foreground regions.

A total of 1000 images from the ASD dataset were used to thoroughly evaluate the performance of the methods. For the modified methods, appropriate maps are first selected by the DFS method before combination using the learned combination scheme. The average MAE scores for original and modified methods are presented in Figure 9.2 (a) and (b). The results suggest that the FGSopt-DFS method successfully removed noisy features to achieve a substantial performance gain of about 12%. The results of a two-sided Wilcoxon ranksum test confirmed that the MAE scores

obtained by the methods are statistically different with a  $p$ -value =  $2.17 \times 10^{-4}$ . It can also be observed that the MAE distribution of FGSopt-DFS is more concentrated at lower MAE values. In the case of IGA, substantial improvements of about 30% are observed after the introduction of DFS. The reason for comparatively large improvements for IGA is that the features employed by IGA are highly noisy on difficult images. Hence the DFS method has more scope for improvement in case of IGA as compared to the FGSopt method. The results of a two-sided Wilcoxon ranksum test suggest that IGA-DFS is highly statistically different from the IGA method as indicated by a  $p$ -value  $< 0.0001$ .

In order to explore a justification for these improvements, specific cases of detection and rejection of outlier features are investigated. Figures 9.3 and 9.4 present example cases where noisy features are identified and neglected and only appropriate features are selected for combination. The process of grouping good and bad features using agglomerative hierarchical clustering is depicted. In Figures 9.3 and 9.4, the height of the “ $\square$ ” shapes depicts the euclidean distance between objects. The green lines signify joints from where the hierarchical tree is cut to separate good and bad features. The blue ellipses indicate features selected by the dynamic feature selection method for combination based on their quality.

Figure 9.3 depicts two scenarios where a set of high quality maps are segregated from the noisy feature maps. It is not difficult to follow that the sorting of features according to their quality makes the partitions obvious and it becomes trivial to split the hierarchical tree into disjoint clusters. These results signify the effectiveness of the feature discrimination property of the DFS system.

The top row example in Figure 9.4 presents a case where every feature excluding the selected features contain background noise that can potentially corrupt the final saliency output. The bottom row presents a special case, where only a single feature is selected as all other features are corrupted with unwanted background information. This result highlights the

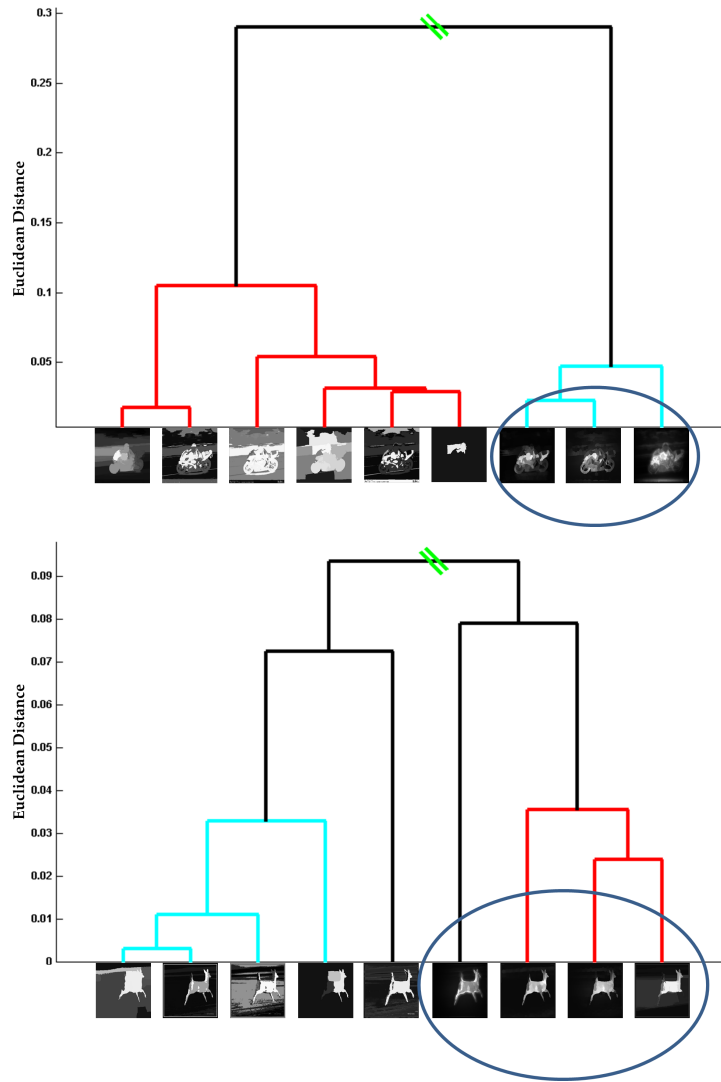


Figure 9.3: Visualisation of dynamic feature selection based on quality dependent hierarchical clustering of the DFS method. Both figures show cases where multiple noisy features are neglected and only the appropriate maps are included in the combination process. Learned feature weightings, normalisation and integration functions are applied only on selected features during combination. The red, blue and black colours are assigned to group of nodes whose distance is less than a threshold. The colour assignment is unique and random. The green lines depict the points where the dendrogram is split into two distinct clusters.

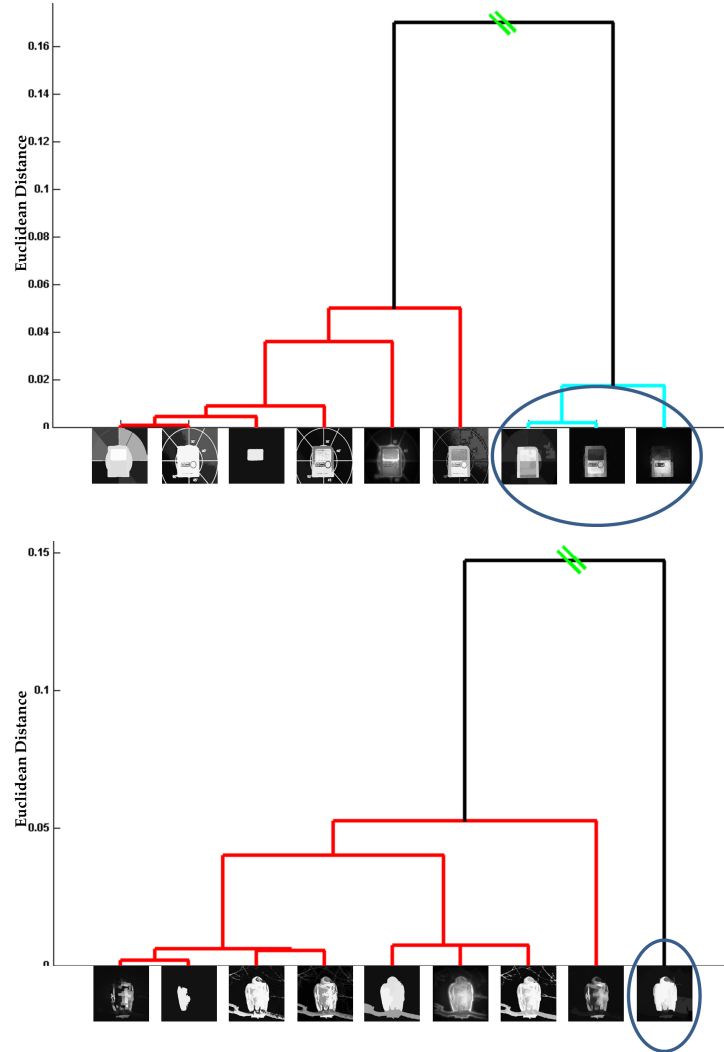


Figure 9.4: Illustrations of the clustering process based on feature quality. The bottom figure presents a special case of DFS where only the best quality feature is selected as all other features potentially induce noise in the final saliency output. The red, blue and black colours are assigned to group of nodes whose distance is less than a threshold. The colour assignment is unique and random.

ability of the DFS method to select the single best map in scenarios where none of the other features is adding any useful information.

These improvements come at the cost of additional computational time required for performing dynamic feature selection before combination. On average 0.97 seconds are required to compute the quality of a feature and about 0.77 seconds are required to compute the foreground approximation and perform hierarchical clustering. The additional computational time required for dynamic feature selection linearly scales with the number of features.

Any additional overhead can be controlled by employing features with low computational overhead. In addition, to reduce the computational requirements, the best performing saliency feature can be employed as the foreground approximation at the cost of decreasing saliency prediction accuracy. Further discussions on saliency prediction accuracy versus computational speed are presented in section 9.4.1.

## 9.2 FGS as Foreground Approximation

A new graph-based foreground approximation  $F_G$  was introduced to assist the proposed feature quality measurement cues in chapter 8. A region based approach was proposed due to its unbiased nature towards any particular feature and to complement the region based state-of-the-art approaches selected for saliency aggregation. Due to region based processing, the proposed foreground approximation, denoted by  $F_G$ , can lead to inaccurate segmentation on difficult images. In contrast, a robust FGS approach was introduced with the ability to suppress background noise in images with cluttered background and induce accurate object segmentations (see chapter 4). The FGS method is now investigated as a foreground approximation in comparison with  $F_G$ .

The FGS approach is incorporated in the DFS based salient object detection method as the foreground approximation to construct a DFS-FGS



method. The performance of the DFS method was compared with the modified DFS-FGS method on images from the ASD dataset. The comparison revealed notable improvements of about 4% and 3% in terms of average AUCPR and F-measure, respectively. These improvements are substantial given the already high detection accuracy of the DFS system.

The two major drawbacks of  $F_G$  are elaborated with representative examples in Figure 9.5. The region based approximation is sensitive to images with distractor objects<sup>1</sup> as shown by the first three examples of Figure 9.5. It can be observed that the objects in the first three rows share spatial position and features with image boundaries as shown by the red rectangles. As the graph based formulation of  $F_G$  rely on boundary nodes for saliency computation, background noise is included in its response, resulting in inaccurate segmentation.

The second artifact of the region based approximation can be observed in the last three rows of Figure 9.5. Due to a different contrast of the boundary regions from the regions in the middle of the image, background regions are falsely highlighted. In contrast, the unbiased FGS approach is accurate in segregating the colours of the foreground and background regions. Thereby, it results in smooth saliency maps with uniform highlighting inside object contours and better background suppression.

This improved performance of the FGS approach as compared with  $F_G$  is reflected by the average AUCPR on the ASD dataset, shown in Figure 9.6 (a). The improvements are further quantified by the performance enhancements on selected difficult images from the ASD dataset. Difficult images are selected by considering those images on which both the approaches result in low AUCPR scores.

The timing comparison for the approaches is presented in Figure 9.7. Figure 9.7 (a) compares the computational time in seconds for both the approaches. It can be clearly observed that  $F_G$  is multiple times faster as

---

<sup>1</sup>Distractor objects in this context refer to objects that share features with the boundary regions.

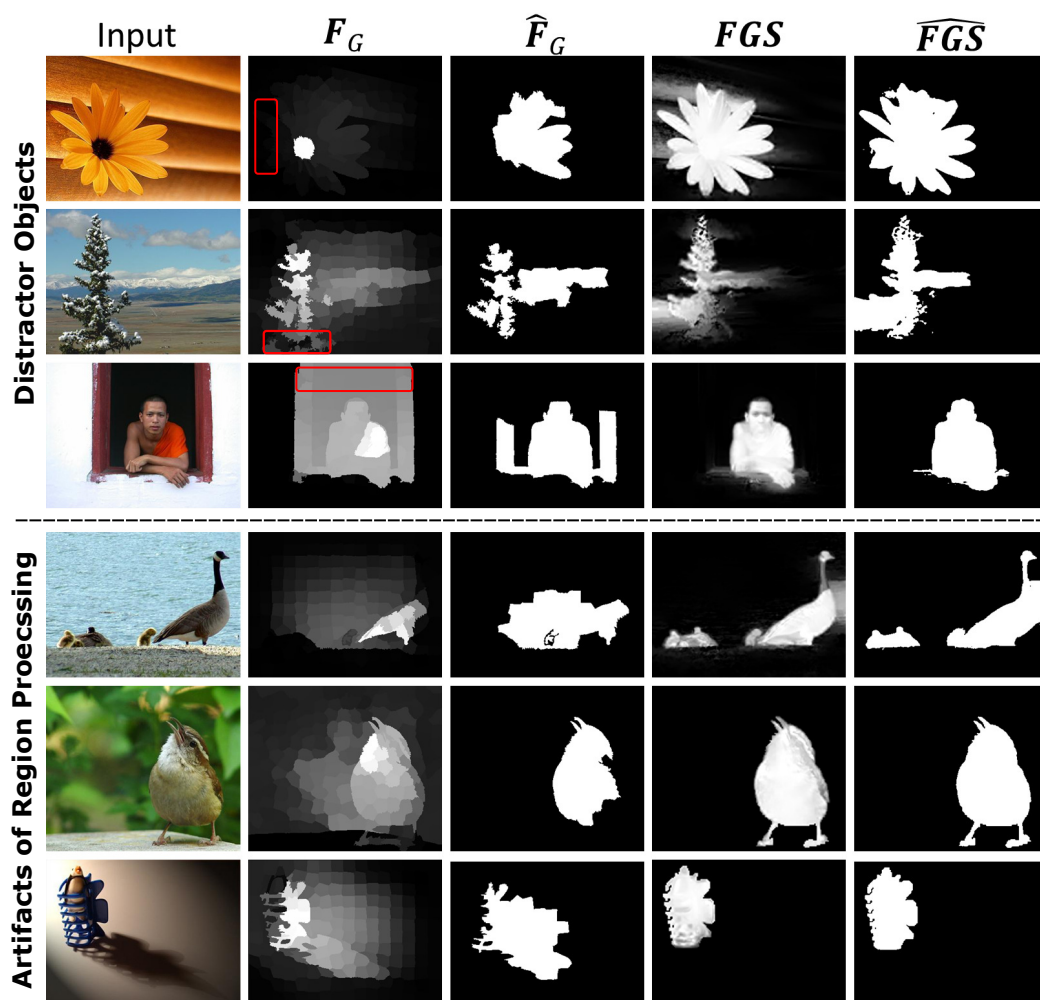


Figure 9.5: Visual comparison of the region based foreground approximation ( $F_G$ ) with the FGS based foreground approximation. From left to right: input images taken from the ASD dataset,  $F_G$ , segmented region based approximation, FGS output and segmented FGS output.

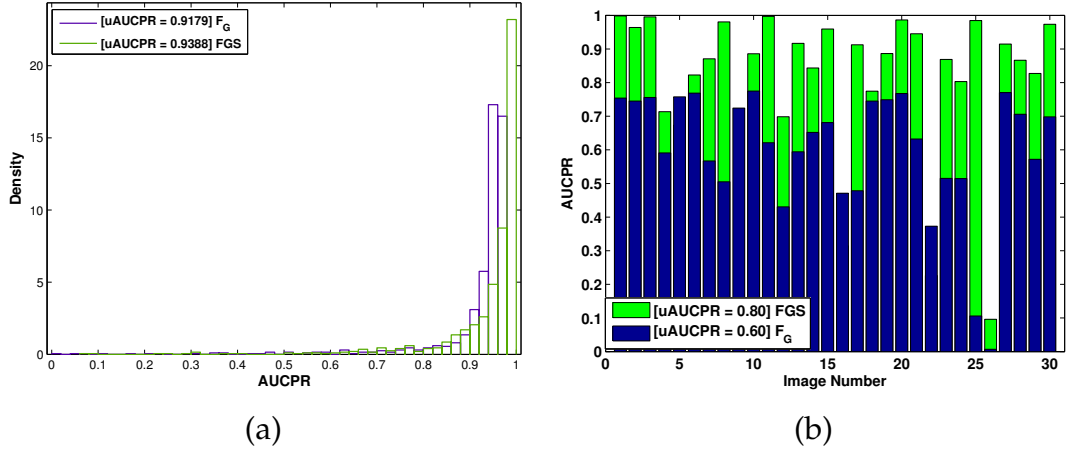


Figure 9.6: (a) Comparison of the foreground approximation approaches in terms of average AUCPR on the ASD dataset. (b) Comparison on selected difficult images from the ASD dataset.

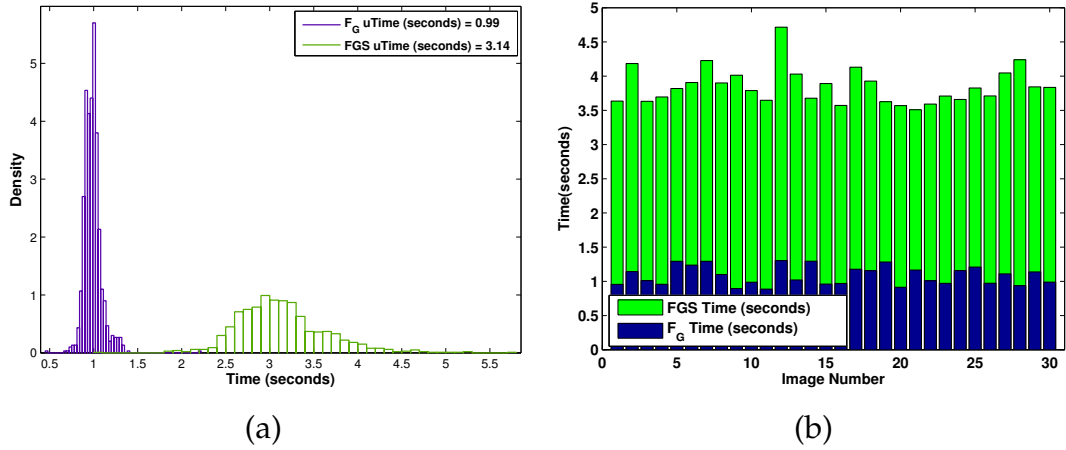


Figure 9.7: (a) Comparison of the foreground approximation approaches based on their computational efficiency. (b) Comparison on selected largest images from the ASD dataset.

compared with the FGS method. This contrast in timing requirements of methods is due to their region versus pixel based processing pipeline. On selected largest images from the ASD dataset, the timing improvements obtained by  $F_G$  are further pronounced.

In many applications, computational time is important, even at the cost of some reduction in accuracy. In most feature selection methods  $F_G$  is therefore a suitable choice for foreground approximation, due to its low computational time without greatly compromising performance. However, in applications where the accuracy of the result is more important than the computational time, such as image cropping and picture collage, the FGS method may be the preferred choice for foreground approximation.

### 9.3 Proposed Techniques Improving State-of-the-art Methods

A variety of methods were introduced to improve the generalisation for salient object detection in this thesis (see chapters 4-8). The effectiveness of these approaches was evaluated both by comparison amongst themselves and by comparison with state-of-the-art methods. Although most of these techniques were introduced to extend and improve previous salient object detection methods, no investigation was done for incorporating them in existing benchmark approaches and recording performance improvements, if any. In this section, two of the proposed techniques in thesis, i.e. FGS and DFS are incorporated in state-of-the-art salient object detection methods to investigate any performance improvements.

### 9.3.1 FGS Improving State-of-the-art Region Based Methods

In this section the, proposed FGS technique is incorporated into three recent region-based benchmark methods. The aim is to investigate whether the smooth and uniform response of the FGS method can aid the benchmark methods to overcome the artifacts of region-based processing. The benchmark methods include DRFI [68], HDCT [76] and UFO [69]. These methods are chosen due to their benchmark performance, recency and region-based processing nature. DRFI employs regional contrast (RC), regional property (RP) and regional backgroundness (RB) features to learn a regressor. On difficult images, the response of DRFI becomes susceptible to background noise due to its regional contrast feature. Hence, the regional contrast feature of DRFI is replaced by the object-aware FGS technique to form DRFI-FGS method. Essentially the response of the FGS method acts as an input feature in the feature processing pipeline of DRFI. HDCT employed a region-based trimap for foreground background approximation. According to the discussion in section 9.2, FGS based foreground approximation was found to be more accurate than a region-based approximation. Therefore, FGS is incorporated into HDCT as its trimap to form HDCT-FGS method. UFO combines the uniqueness ( $\mathcal{U}$ ), focusness ( $\mathcal{F}$ ) and objectness ( $\mathcal{O}$ ) of regions to construct its output response. As the FGS method captures accurate object saliency, it is employed in place of the objectness feature in UFO to construct the UFO-FGS method.

The performance of the original and FGS based methods are compared using four different measures including F-max, mean absolute error (MAE),  $F_\beta$  according to the adaptive thresholding benchmark [3] and average precision (AP). F-max and AP evaluate the robustness of the saliency response to fixed thresholding [3],  $F_\beta$  evaluates the saliency response for adaptive threshold during region based segmentation, while the MAE measure evaluates how correctly the saliency output predicts the background regions.

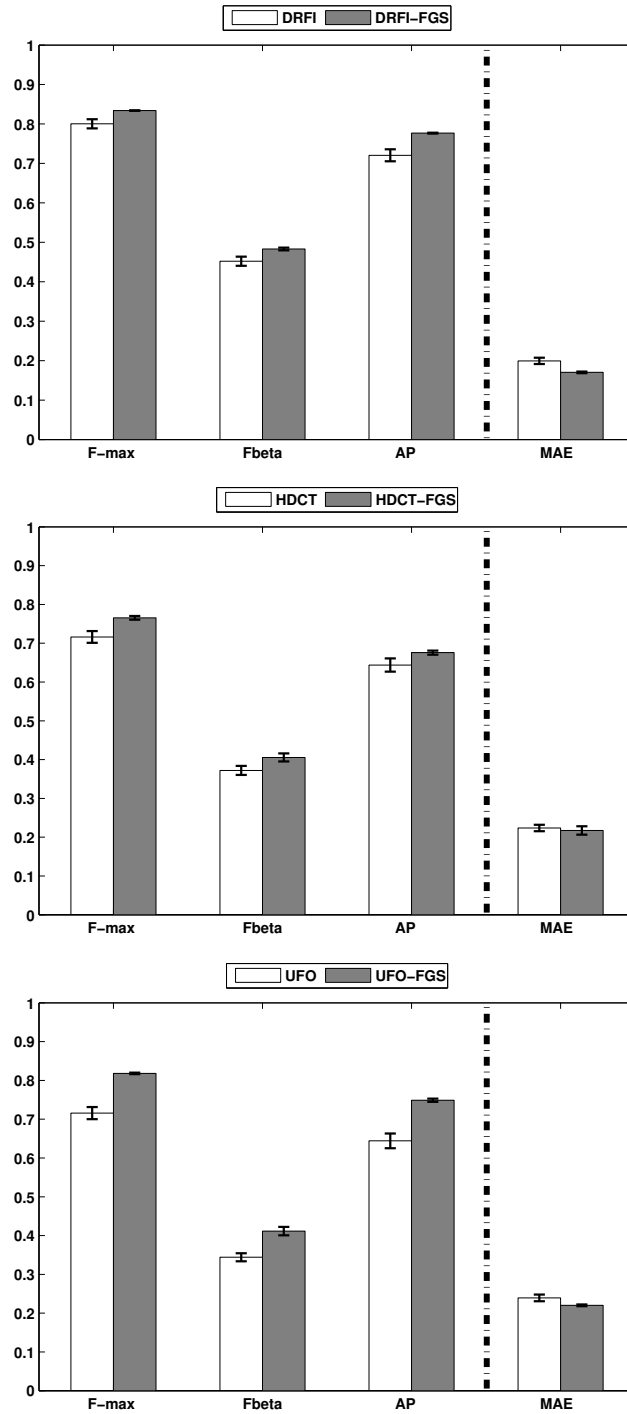


Figure 9.8: Performance comparisons of selected methods with their FGS based improved versions using images from the BSD dataset. From top to bottom: DRFI versus DRFI-FGS, HDCT versus HDCT-FGS and UFO versus UFO-FGS. The errorbars present the standard error of the mean. Four performance measures including maximum F-measure (F-max),  $F_{\beta}$ , average precision (AP) and mean absolute error (MAE) are employed. Large values of F-max,  $F_{\beta}$  and AP are desirable, while MAE (separated by the dashed line in the plot) is to be minimised.

Figure 9.8 shows the performance comparison of the original methods with the FGS incorporated modified methods on images from the BSD dataset. BSD is chosen for this evaluation as it contains a wide variety of image classes with high detection difficulty. The top figure compares the average performance of the DRFI method with the modified DRFI-FGS method. The DRFI-FGS method improved upon the DRFI method with noteworthy improvements of 4.1%, 6.8% and 7.8% in terms of F-max,  $F_\beta$  and AP measures, respectively. This improvement in the segmentation performance of DRFI can be attributed to the smooth and uniform response of the FGS method. A more pronounced improvement of 14.5% was observed in the background prediction capability of DRFI. This result suggests that the noise suppression property of the FGS method helped the DRFI method in discriminating foreground regions from background regions on the noisy images of the BSD dataset. Notably, the saliency prediction confidence of the DRFI method is also improved as depicted by the lower deviation of the DRFI-FGS method from the mean performance measures.

The middle plot shows the average performance of the HDCT approach with the modified HDCT-FGS method. Nominal improvement of 2.8% achieved in terms of MAE suggest that FGS does not substantially improve the background prediction capability of HDCT. The reason for this result is that despite the accurate prediction of the foreground and background regions by FGS, the region based features of the colour spaces already include background noise, which appears in the final response. Substantial improvements of about 6.8% and 8.9% in terms of F-max and  $F_\beta$  suggest that FGS helped the HDCT method to highlight more regions inside object and assign them uniform saliency.

Notable improvements of greater than 10% were observed in terms of F-max,  $F_\beta$  and AP for UFO as shown in the bottom plot of Figure 9.8. These improvements suggest that FGS improved the response of UFO by replacing the region-based non-uniform saliency response with smooth

uniform saliency inside object contours.

Two-tailed t-tests were performed to investigate the statistical significance of the measures after confirming the normal distribution assumptions. The results from the original and modified methods were found to be highly statistically significant with p-values  $< 0.0001$  for all the comparisons.

Figure 9.9 presents the visual comparison of the original and FGS incorporated benchmark methods on representative instances from the BSD dataset. For DRFI and UFO, the input, ground truth, FGS response, original and modified outputs are shown. The visual comparison for HDCT also includes the trimaps generated by HDCT and FGS approaches. The quantitative performance of the images presented in Figure 9.9 is shown by Figure 9.10 with column 1 and column 2 depicting performance for Image 1 and Image 2, respectively.

In the case of DRFI, the first image presents a scenario where background noise is included in the response of DRFI. The proposed FGS approach helps DRFI to discriminate between foreground and background making the salient objects more prominent and smoothly highlighted. The effect of uniform saliency assignment obtained by the DRFI-FGS method can be observed by the segmentation quality measures for Image 1 as depicted by row one and column one of Figure 9.10. The noteworthy background suppression achieved is also evident by the decrease in respective MAE measure in the same plot. The second representative image profiles the characteristic region based saliency assignment response of the DRFI method. The FGS method helps DRFI by promoting similar saliency values across different regions of the salient object and uniform saliency assignment inside object contours. The result is reflected by better region-wise segmentation quality as depicted by the enhanced  $F_\beta$  measure in column two, row one of Figure 9.10.

For Image 1, Figure 9.10, the response of HDCT is corrupted by the partial occlusion due to the bushes. It is an artifact of HDCT's region based



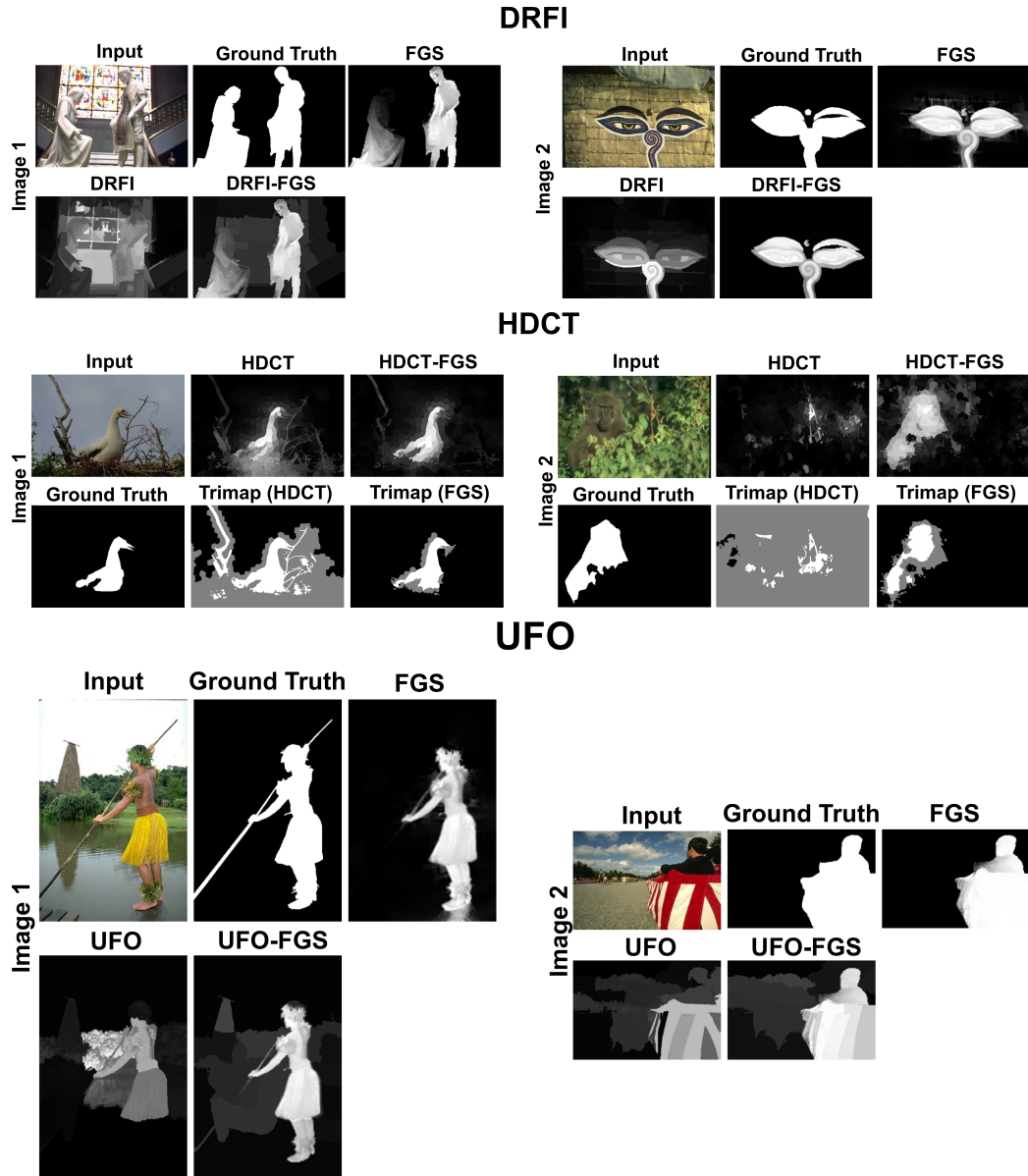


Figure 9.9: Comparison of the original and modified methods on representative images from the BSD dataset.

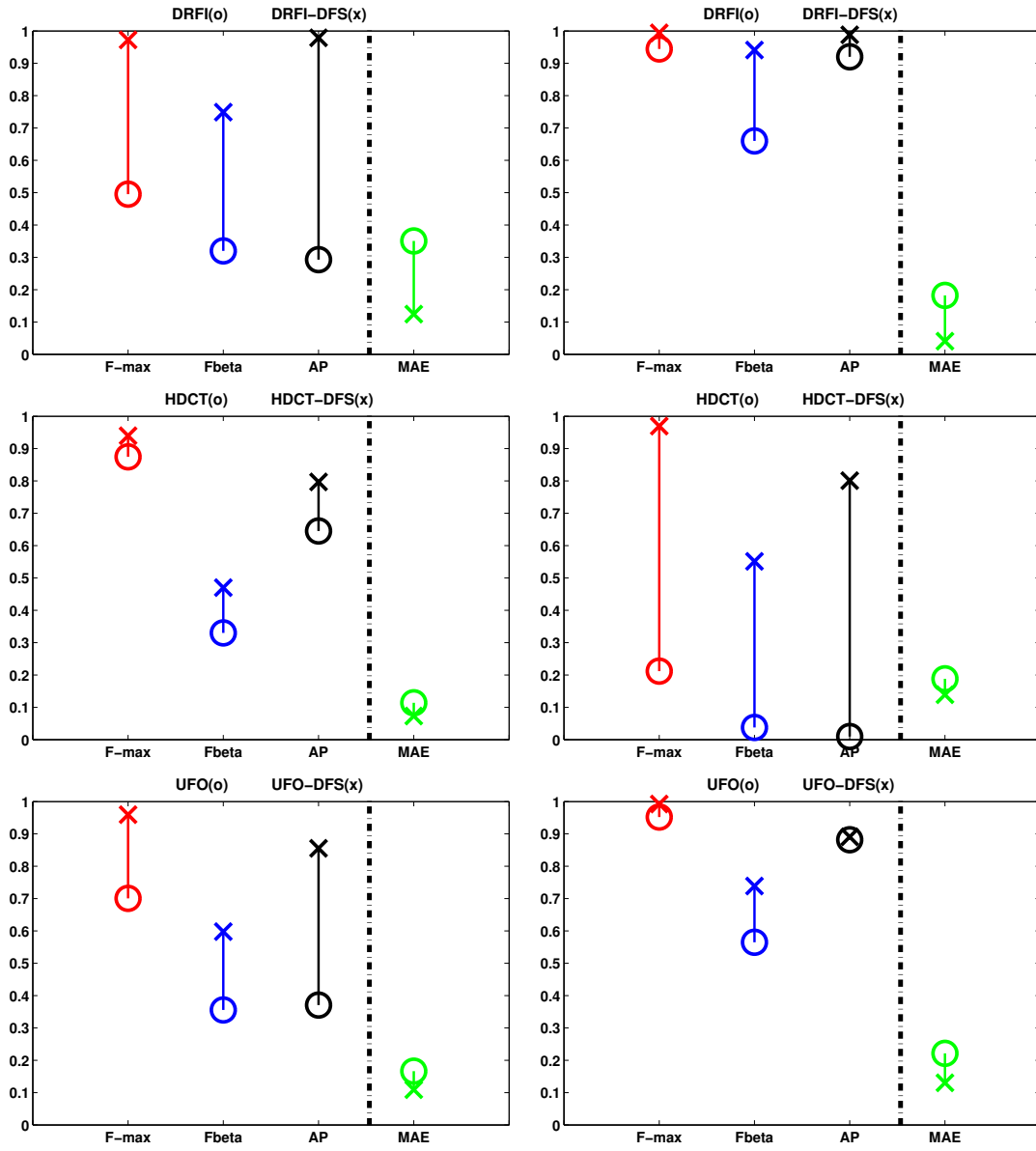


Figure 9.10: The columns present the quantitative results for Image 1 and Image 2 respectively for each method. The lines between the (x) and (o) depict the respective improvements achieved by the modified FGS based methods as compared with the original methods. F-max,  $F_\beta$  and AP are maximised, while MAE separated by the line in the plot is to be minimised.

trimap that labels part of the bush as definite foreground. In contrast, the less noisy FGS based trimap better assists the combination of colour spaces for HDCT-FGS, thereby improving upon HDCT. Image 2 presents an example, where the trimap of HDCT completely fails in segregating true foreground and background. The reason for this result is that the trimap of HDCT mainly considers colour contrast of regions to discriminate foreground and background regions. As there exists low contrast between the salient object and part of the bush, it samples distinct regions from the salient object and the bush as definite foreground in the trimap. In contrast, the FGS method separates foreground and background colours globally in eigenvector space, thus it is able to find partitions between the salient object and the background colours. The result is reflected by the accurate trimap constructed based on the FGS method.

Finally, for UFO both Image 1 and Image 2 present cases where non-uniform region-wise saliency assignment is observed in the UFO response. The FGS method improves the UFO response by uniform assignment of saliency inside object contours. A pronounced example of this scenario can be observed by the saliency assignment to the persons head and body in the respective responses of UFO and UFO-FGS methods in Image 2.

### 9.3.2 DFS Improving Complementary Feature Selection Methods

The proposed DFS method was employed for saliency aggregation for salient object detection and notable improvements over state-of-the-art methods were reported (see chapter 8). However, no investigation was performed to evaluate its effectiveness for complementary salient feature selection as attempted by methods in literature. In this section two experiments are performed that employ the proposed DFS method for the task of complementary feature selection. Experiment I compares the feature selection performance of the proposed DFS method with the com-

pactness (SV) measure of Cheng et al. [27] for selection of the complementary global saliency cues of [27] (i.e. colour spatial distribution (CSD) and global uniqueness (GU)). In experiment II, the feature selection performance of the proposed DFS method is compared against the SI approach of Gopalakrishnan et al. [49] (termed as SIsal) using their complementary saliency features, i.e. colour saliency framework (CSF) and orientation saliency framework (OSF). Details of the methods used for comparison are reported in chapter 2.

### Experiment I

To obtain the global saliency cues of colour spatial distribution (CSD) and global uniqueness (GU), the MATLAB implementation provided by the authors [27] is used. The GC method of [27] uses the compactness measure of [49] for feature selection, which is implemented in MATLAB for this work. The performance comparison of the proposed DFS method with GC is shown in Figure 9.11.

In the top row of Figure 9.11, the performance of the proposed DFS shows slight improvement over GC on the MSRA10k dataset. It is anticipated that this is due to the output of the individual features, which is highly complementary in the case of MSRA10k and selecting the best feature is not a very difficult task. Hence DFS is able to produce only modest improvements. Our hypothesis is supported by the noteworthy improvement obtained by the proposed method over the GC method on the more difficult BSD and SED2 datasets. The proposed DFS method shows improved performance on all thresholds with higher F-max on BSD and SED2 datasets. It is to be noted that GU's performance on the SED2 dataset is substantially better than GC, due to the false feature selections made by the GC method on the SED2 images.

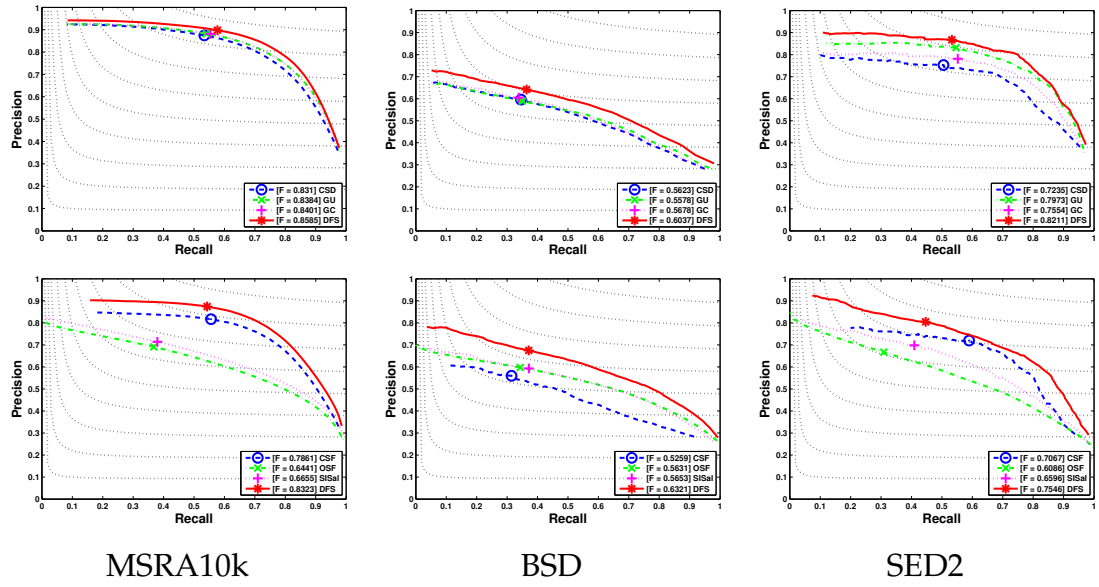


Figure 9.11: Top row (left to right): precision-recall curves for experiment I on MSRA10k, BSD and SED2 datasets, respectively. Bottom row (left to right): same set of results for experiment II. The proposed DFS method is shown in red. Isolines on the plots show the contour for F-measure in the range 0.1-0.9 in steps of 0.1.

## Experiment II

The goal of the second experiment is to select the best feature from the complementary features, i.e. colour saliency framework (CSF) and orientation saliency framework (OSF). The complementary nature of these features is faithfully captured by implementing them according to the author described settings [49]. The quantitative comparison results are presented in the bottom row of Figure 9.11. It can be observed that the proposed DFS method improves upon the performance of SISal on all three datasets and on all thresholds. The DFS performance is substantially higher than SI based selection scheme on these features as compared to the previous experiment. This can be explained by the observation that the CSF feature is less compact in most cases, but captures the salient object better than OSF, resulting in high false selection rate by SISal. In contrast the proposed DFS method is focused towards measuring the figure-ground proposal quality, therefore selecting better performing features. It is also to be noted that CSF has higher precision than SISal on both MSRA10k and SED2 on similar recall values.

Figure 9.12 presents the visual comparison of the proposed DFS approach with GC and SISal methods, respectively. The first and second row images are taken from the MSRA10k dataset, the third and fourth row images are taken from the BSD dataset, and the last row image is taken from the SED2 dataset. It is to be noted that SISal favours the OSF feature for all images, despite its inferior salient object prediction capability. The reason for this response is the compactness and connectivity of the OSF maps, which is preferred by SISal at the cost of low salient object detection and segmentation ability. Conversely, the spatial variance of the GC method selects features having good figure-ground proposals for the third and fourth row image. However less spatial variance in the horizontal and vertical directions influences the GC method to select the OSF map for all other images. The proposed DFS method in contrast uses its foreground approximation and feature quality cues to select features that provide bet-

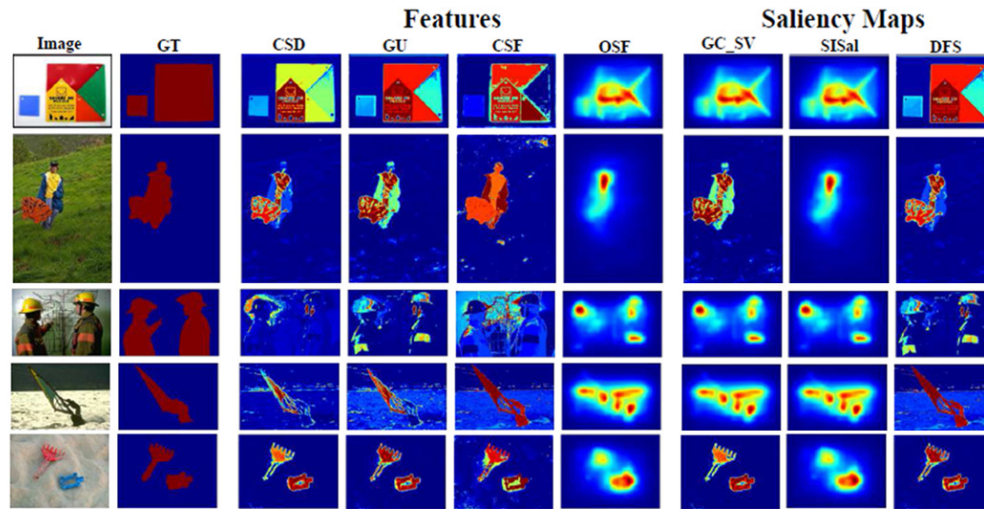


Figure 9.12: Visual comparison of the proposed DFS method with the GC and SISal methods on representative images from MSRA10k, BSD and SED2 datasets. From left to right: input, ground truth (GT), features of the GC and SISal methods and the respective saliency maps for GC, SISal and the proposed method after complementary feature selection.

ter figure-ground proposals. It can be observed that features selected by the DFS method have higher density inside predicted salient objects, better reconstruction capability, ability to induce better segmentation quality and also higher edge energy.

## 9.4 Recommendations for Incorporating the Proposed Techniques For Improved Salient Object Detection

The following are a set of guidelines for current and future salient object detection methods for improving their generalisation performance on unseen difficult images based on the findings of this work. Figure 9.13 elaborates these guidelines. The diagram provides a simple version of the effects of various contributions on the overall system. For descriptive purposes, the dynamic feature selection and learned feature combination are depicted as isolated processes. Other methods can employ the required variant of the dynamic feature selection method and choose a learned combination method for subsequent use. The choice of a feature selection variant and feature combination method depends upon the required performance metric, i.e. generalisation/robustness versus computational time and training overhead. A more holistic view of the effects of the contributions on the overall system and their interaction is provided in the text.

### 9.4.1 Recommendations for Incorporating Dynamic feature Selection

Based on the results obtained in sections 9.1 and 9.3.2, it can be deduced that DFS is beneficial for improving generalisation performance on unseen



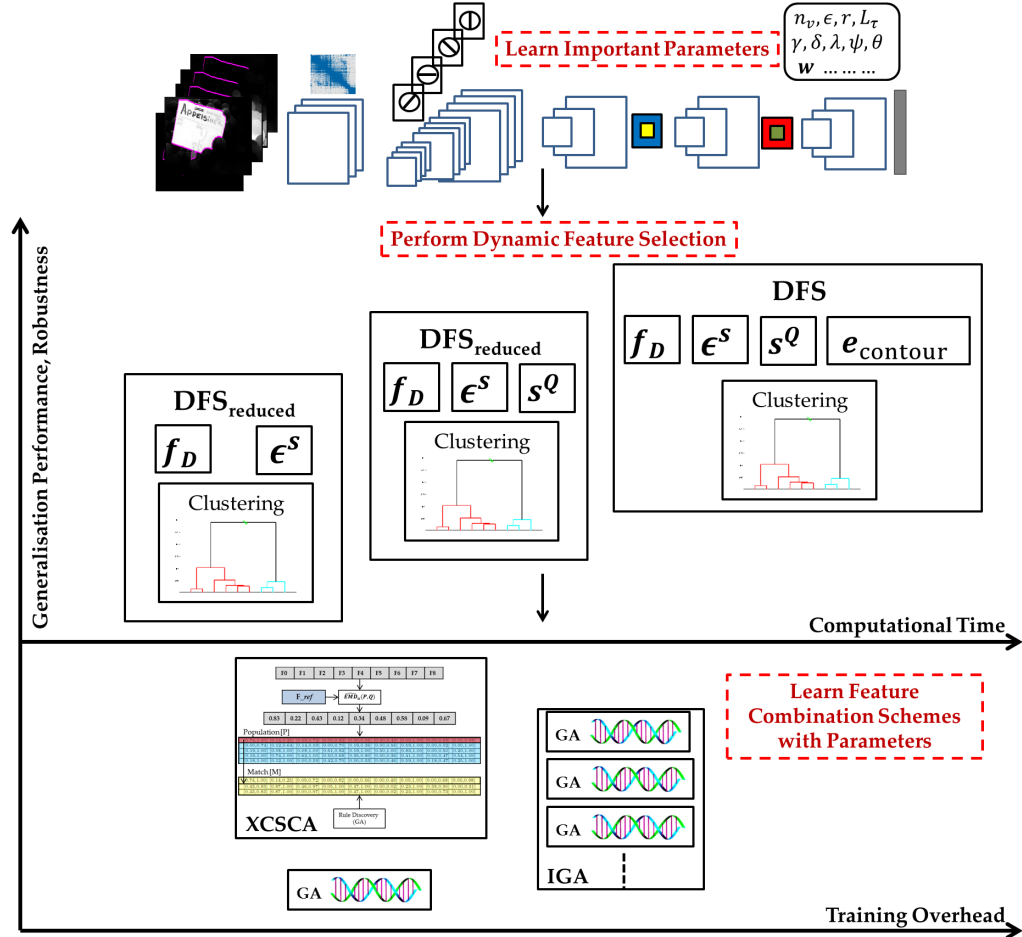


Figure 9.13: Guidelines for current and future salient object detection methods for incorporating the methods proposed in this thesis. The generalisation versus time/overhead graph is not drawn according to scale and the measures on the x-and y-axis are only for illustration.

images. However, there is a trade-off between improved saliency prediction accuracy and increased computational time.

For methods that require increased robustness and improved generalisation for non-real time applications, the DFS method can act as an off-the-shelf method that can be readily incorporated into the feature processing pipeline. For further improvements in saliency prediction accuracy for difficult unseen images, it is advisable to use the proposed FGS method as the foreground approximation.

Conversely methods that require modest improvements in generalisation without significant computational overhead are recommended to employ reduced variants of the DFS method as discussed below.

The additional computational overhead imposed by the DFS method can be divided into the time required to compute feature quality measurement cues and the computation time for the foreground approximation. To reduce the overhead for foreground approximation computation at the cost of compromising accuracy, a feature with good performance and low computational overhead may be employed as the foreground approximation.

To reduce the computational overhead of the feature quality cues, a subset of cues depending upon the application can be employed. The cues  $f_D$  and  $\epsilon^s$  judge a feature map's ability to uniformly highlight a saliency object and can be regarded as uniform saliency assignment cues. While the  $s_Q$  and  $e_{\text{contour}}$  evaluate the segmentation quality of a feature map and hence can be regarded as the segmentation quality cues. Therefore, if the salient object method is designed for applications that require the result to be segmented at a variety of thresholds, the uniform saliency assignment cues should be applied in isolation. Whereas, if the target application requires the output to induce quality segmentations at certain specified thresholds, the segmentation cues may be used separately.

Furthermore, the per feature computation time required by the feature quality cues is highly dominated by  $e_{\text{contour}}$  as the time required by the

other cues is merely a tenth of a second. The  $e_{\text{contour}}$  cue is included in the DFS method to provide completeness in covering the cue space and provides small additional improvements as supported by the results in section 9.3.2. Therefore, for significant reduction in additional computational time at the cost of insignificant loss in performance, the  $e_{\text{contour}}$  cue should not be included by the methods that require fast processing.

## 9.4.2 Recommendations for Incorporating Learning Methods

### Comparison of the FGS and FGSopt methods

The FGSopt method improved the salient object detection performance of the FGS method making it competitive with the best performing benchmark methods. An improvement over FGS of approximately 4% was achieved in terms of area under the PR curve. This improvement was obtained at the cost of additional computational time at the training and test stages. With reference to the computational time at the test stage, the eigenvector features do not add any time as they have to be computed to obtain the matting components in both methods. However, the CSD features add more time at the test stage. For instance, for an image of size  $300 \times 400$  it increases the average time from  $2.52 \pm 0.23$ s for FGS to around  $7.30 \pm 1.29$ s for FGSopt. This is still practical as the average time required by DRFI (being the most robust model in terms of quantitative performance) in our experiments for the same images is around  $13.23 \pm 1.31$ s<sup>2</sup>. The processing time of the FGSopt method is also better than the two recent benchmark methods, LRK [123] and BSM [144].

---

<sup>2</sup>The timings reported here were computed using an i7 vpro 3.2 GHz processor with 8 GB of RAM for 100 randomly sampled images from the MSRA dataset. The same images and system settings are employed to compute all the timings reported in this section.

### Comparison of Joint Optimisation and the Multiple Combination Schemes Based Methods

The formulation for joint optimisation introduced in chapter 5 is in principle different from the multiple combination schemes based methods discussed in chapters 6 and 7. The joint optimisation based single combination scheme is meant for methods that attempt to combine complementary features that are not highly varying in nature, and where it is not highly challenging to assign relative importance to them. The significance of this combination scheme relies in finding the relationship between variables and the features while combining them optimally. The latter task requires extensive search to find the best single solution that fits the problem. This is in essence different from the multiple combination schemes based methods, which are designed for features that are highly varying in their performance and depend on the type of image operated upon. These methods naturally require multiple solutions, each fitting a particular performance landscape of features. Therefore, though it is expected that the multiple combination schemes might benefit FGSopt, it is not recommended to employ these approaches to a joint optimisation problem, as discussed in chapter 5.

In addition to the problem at hand, other differences that drive the choice of the learning schemes are the training procedures and the computational time at the test stage. In general at the training stage, the prime difference between single and multiple combination schemes is the generalisation performance versus the training overhead. Specifically the computational complexities of the objective functions at the training stage for the single and multiple combination schemes are dependent upon the respective problems they attempt to solve. For instance, the objective function employed for the single scheme in chapter 5 has a computational complexity of  $O(4n\zeta)$ , while the complexity of the objective function introduced in chapter 6 is  $O(4n)$ , i.e. (6.5). Here  $n$  is the number of pixels of a saliency map and  $\zeta$  represents the number of thresholds employed in

the objective function of chapter 5, i.e. (5.1). The difference in the computational complexity of the objective functions is due to the different problems they attempt to address. For the joint optimisation, extensive search is required to find the best fit solution. Hence accurate ranking of salient object results is required to compute accurate fitness. Therefore, multiple segmentation thresholds in the form of  $\zeta$  are required to accurately rank salient object detection results. In contrast, the objective function for the multiple scheme based method does not require extensive search as the best solution tailored to a single image type (in a niche) is required. Hence, the accuracy of the objective function in judging a saliency map is not as important as in the former case. Therefore, a single segmentation threshold is employed, which implies  $\zeta=1$ . Accordingly the complexity of the objective function in (6.5) is less than the objective function of (5.1).

At the test stage, the FGSopt method spends less time on average as compared with the computational times spent by the multiple schemes based methods (for details see the next section). This is due to the FGSopt method only requiring feature computation time at the test stage in contrast to the multiple schemes based methods, which require additional computational times as discussed in the next section.

### **Comparison of the IGA and XCSCA methods**

When comparing the IGA and XCSCA methods, XCSCA obtained substantial improvements of 3.5% in terms of AUCPR around 5.6% in terms of the classification accuracy. Additionally, the XCSCA method has better generalisability as compared with the IGA method depicted by its better identification of image type.

In terms of computational time at the test stage, on average the IGA method requires  $12.00 \pm 3.09$ s to process 100 randomly sampled images of size  $300 \times 400$  from the MSRA dataset. Whereas, on average the XCSCA method requires  $14.04 \pm 3.38$ s for the same set of images. The times spent by the IGA and XCS methods on an image at the test stage are quite subjec-

tive. As both the methods employ the same set of features, the time spent for feature computation is the same. For the IGA method, the additional time is spent in normalisation and integration schemes being applied. As the normalisation and integration schemes can be quite different for each image (depending upon image type with no normalisation and summation being one of the option), the time spent at test stage is subjective<sup>3</sup>. In case of the XCSCA method, for each test input the classification accuracy is computed by the current action for all the classifiers in the match set and the classifier with the maximum accuracy is selected. Hence, the additional time at the test stage can be varying depending upon the classifiers in the match set that support the action according to the input image type.

The XCSCA method, despite of its additional computational time at the test stage still would be the preferred choice for methods that require multiple combination rules. However, searching for the appropriate normalisation and integration schemes is not inherently best suited to an XCS based method, as it is a niche based classification method rather than an optimisation technique. Genetic Algorithms in contrast are competitive in searching for such variables that are literally encoded as integers in the chromosome. Therefore, for methods that require to search for the best suited normalisation and integration scheme along with the feature importance weights, the IGA method may be the preferred choice.

### **FGS in Multiple Combination Schemes Based Methods**

Following from the above discussion, a related question is that how will the FGS method affect the multiple combination schemes methods, if employed as a feature. The performance comparison of the FGS method with the best performing feature of the multiple scheme methods, reveal performance improvements of around 9% in terms of area under the precision-

---

<sup>3</sup>As the normalisation and integration schemes are quite simple, they do not add a substantial amount of computational time.

recall curve. These improvements are highly likely to substantially improve the overall performance of the multiple learner methods. However, incorporating FGS as a feature would not greatly affect the generalisation performance of the methods to unseen images. This is because it is the learned combination schemes and the ability to accurately identify an image type, which increase the generality of the methods. In terms of computational time, the average time required to process an image of size  $300 \times 400$  for the FGS method is  $2.52 \pm 0.23$ s. This is less than the average time required to compute the most computationally expensive feature of the multiple learner methods, i.e.  $3.30 \pm 0.37$ s. Notably, the computational overhead added by the FGS method can be reduced by discarding a few low performance features.





# Chapter 10

## Conclusions and Future Work

The overall goal of this thesis was to increase the generalisation of the traditional computational model of visual attention by utilising machine learning methods to select and optimally combine appropriate features. This goal was achieved by devising a number of new methods for learning feature combination and feature selection before combination. The developed methods were evaluated using standard benchmarks for salient object detection on benchmark datasets and extensive comparisons were performed with state-of-the-art saliency methods. Where it was observed that no features exist that can completely isolate the foreground regions from the background, the matting based feature was introduced, which suppresses the background information to effectively segregate foreground regions.

The rest of this chapter presents the achieved objectives, the main conclusions deduced from the results chapters and finally points out future research directions that originate from this work.

### 10.1 Achieved Objectives

The following research objectives have been accomplished to achieve the overall research goal.

1. Joint optimisation of feature related parameters and feature importance weights was introduced to improve the saliency detection performance of the traditional visual attention model. Important parameters of the feature computation process and feature importance weights were learned by optimising a task specific objective function for human fixation prediction. By maximising the agreement between predicted saliency and the target (human fixations), the proposed GAOVSM method improved upon the performance of eight deterministic state-of-the-art saliency detection techniques on the task of human fixation prediction.
2. Spectral matting was employed for the first time in saliency prediction to combat the artifacts of region-based approaches for salient object detection. A novel saliency method for figure-ground segregation (termed as FGS) was introduced that employed matting components to construct smooth, uniform and accurate saliency maps. The FGS method was able to overcome the artifacts of regions-based approaches by assigning uniform saliency to object regions while suppressing unwanted background information. As supported by the quantitative performance on salient object detection, the proposed FGS approach improved upon several state-of-the-art techniques.
3. Joint optimisation of feature computation parameters and feature importance weights was introduced for optimal combination of FGS with complementary features. Feature related parameters and their respective importance was learned at multiple segmentation thresholds by maximising the area under the precision-recall curve as an objective. The developed FGSopt method improved the object detection performance of the FGS technique and improved upon several state-of-the-art salient object detection models by considering the performance gaps amongst features.
4. *Semi-Autonomous identification of image type* was introduced to learn

multiple feature combination schemes. Multiple combination schemes were learned from distinct image groups using multiple Genetic Algorithms (GAs). Images were placed into distinct groups using a semi-autonomous method that relied on their feature composition. By employing a suitable combination scheme for each unseen image type, the proposed image based GA (IGA) approach exhibited better generalisation as compared with a baseline GA that learned a single combination scheme. The IGA method also exhibited significantly better performance as compared with two classification based benchmark methods and three state-of-the-art models on the task of salient object detection.

5. Introduced *autonomous identification of image types* for learning multiple feature importance rules in order to increase the generalisability of the system on unseen image types. A supervised XCS based method was introduced that divided the search space into niches in order to learn effective feature importance rules. This was achieved by employing a novel encoding scheme and a suitable action computation function. The proposed XCS based method improved upon the performance of the previously proposed multiple GA based method by obtaining a set of generalised feature importance rules.
6. Novel cues were established for dynamic feature selection in order to advance the current state of complementary feature selection and feature/saliency aggregation. Saliency quality measuring cues were introduced to seek discriminative features for appropriate combinations. Label free measurement of feature quality enabled the proposed feature selection method to improve upon the state-of-the-art in complementary feature selection and saliency aggregation. The proposed DFS based object detection method also improved upon seven state-of-the-art salient object detection methods.

## 10.2 Conclusions

### 10.2.1 Joint Optimisation of Important Parameters and Feature Importance

#### Joint Learning for the Traditional Model of Attention

Joint learning of feature related parameters and feature importance weights was introduced in the traditional model of attention (see chapter 3). The goal of learning important parameters of the traditional attention model was successfully achieved by Genetic Algorithm based search for important parameters by the introduced GAOVSM method. Learning suitable values for important parameters of the feature computation process yielded substantial feature level performance improvements of 22.5%, 77% and 50% in terms of average AUC,  $S$  and NSS measures. Moreover, the 99% confidence intervals confirmed that the proposed optimised features always improve the performance of the traditional model of attention. This is an important result as it demonstrates the importance of tuning parameters of the traditional model of visual attention. Additionally the visual comparison of the unoptimised and optimised feature set revealed that the optimised version better matches the desired ground truth at pixel level, as compared with the baseline model.

Analysis of evolved solutions revealed important findings. Confirming previous studies [17], the optimised parameters resulted in narrow bandwidth filters tuned to object background discrimination in cluttered scenes. Also the receptive field was tuned to elongated contours of objects. The optimised orientations exhibited bias for vertical orientations following the human visual cortex [89].

When the method with optimised parameters, i.e. GAOVSM was employed, notable improvements were obtained at the test stage as compared with the unoptimised method. In agreement with prior studies [153], the orientation feature was found to be heavily weighted as compared to the

colour and intensity features. The improved orientation feature in coupling with its learned higher importance enabled the GAOVSM method to generalise better on unseen images.

GAOVSM outperformed several benchmark methods in terms of AUC and  $S$  score for unseen test images. In terms of the NSS measure, RARE performs better than other methods including GAOVSM due to its extra normalisation step on the saliency output. However, GAOVSM exhibited more confidence and lower deviation from the mean in terms of NSS as compared with RARE and showed comparable performance even with the lack of a normalisation step.

### **Introduction of Figure-ground Segregation Based Feature**

Despite the promising performance of the joint optimisation based traditional model on fixation prediction, it was not suitable for salient object detection due to its simplistic features. The features employed by the benchmark methods produced undesired artifacts due to their region-based processing. Therefore, a new matting based feature was introduced to combat the artifacts of region-based processing. The feature was computed by selecting only those matting components that are predicted to belong to the salient object by the figure-ground segregation (FGS) method introduced in chapter 4. The proposed FGS method overcame the problems introduced by the region-based processing methods by uniformly highlighting complete salient objects and effectively suppressing background noise. The FGS method was able to improve upon several benchmark methods with an improvement of 2.6% as compared with the best performing method amongst them.

With the help of examples it was shown that the recorded performance improvements were due to the good object coverage of the matting components and the capacity of the FGS method to reject unwanted noisy matting components.

### **Improving FGS by Joint Learning of Parameters**

A Genetic Algorithm based joint optimisation approach for optimal combination of FGS with complementary features was investigated in chapter 5. The goal of searching for feature related parameters along with generalised feature importance was successfully achieved by minimization of the difference between predicted saliency and ground truth annotations.

The results of the FGSopt method confirmed that the complementary features aid FGS in rare circumstances preventing loss in performance.

The performance of the optimised eigenvector features demonstrated that they suppress background noise in contrast to the baseline eigenvectors, due to their ability to partition the data into foreground and background clusters. The optimised eigenvectors also enabled the foreground object saliency feature to more closely follow the ground truth profile as compared with its unoptimised version.

The results of the FGSopt method demonstrated that it was able to find suitable parameters and learn an appropriate importance rule to minimize the performance gaps amongst individual features. FGSopt exhibited improvements of 3.8% over the FGS method and 16.2% over the unoptimised method. The 99% confidence intervals on multiple runs of the FGSopt method demonstrated that it always outperformed FGS and the unoptimised version.

The validity and effectiveness of the FGSopt method was further confirmed by the elevated performance of the FGS method up to a competitive degree with the best state-of-the-art methods in terms of both fixed and adaptive thresholding benchmarks. The FGSopt method also outperformed several state-of-the-art methods, exhibiting substantial improvements.

The FGSopt method improves upon the FGS method in terms of quantitative and qualitative performance. These improvements are achieved at the cost of an overhead in training and test time. The comparison between FGS and FGSopt methods is a trade-off between added accuracy at the

cost of additional computation time. The FGSopt method still requires less computational time than DRFI (a top benchmark method) , while achieving competitive quantitative and better qualitative performance.

### 10.2.2 Multiple Combination Schemes Tailored to Image Type

#### Image Based Genetic Algorithm (IGA)

Learning of multiple combination schemes each suited to a particular image type was investigated by employing multiple Genetic Algorithms (see chapter 6). A variety of normalisation and integration techniques reported in past works were investigated to form unique combination schemes. The goal of finding and applying a suitable combination scheme (depending upon image type) was successfully achieved by semi-autonomous grouping of images and genetic search for multiple combination schemes by the IGA method.

It was shown through analysis of the optimised solutions of the proposed method that the feature combinations learned by IGA adjust according to image types. This enables the IGA method to achieve better generalisation than the fixed baseline combination scheme. The IGA method also exhibited better quantitative and qualitative performance than the SVM based benchmark methods employed for salient object detection by prior works. Specifically, AUCPR improvements of 9.8% and 9.3% were observed over the SVM and the NLSVM methods, while noteworthy improvements of 41.5% and 36.6% were recorded in terms of AUF.

The analysis of evolved solutions revealed that the favoured integration type was summation, suggesting appropriate feature conditioning before integration. Iterative and global normalisation schemes were found to cover images with cluttered background and with distractor objects, respectively.

The optimisation framework of the IGA method is designed for com-

bination of complementary features that have high degree of variability in performance on unique image types. In comparison to the single combination rule of the joint optimisation framework, it can offer improved generalisation performance at the cost of added computational overhead at the training stage and computational time at the test stage. When compared with the XCS based method, it has the added advantage of encoding various normalisation and integration approaches in its search formulation.

### **Learning Classifier System with Computed Action**

An XCS based method was investigated for learning multiple feature combination rules termed as XCSCA (see chapter 7). The goal of applying a combination rule according to image type was successfully achieved by autonomous identification of image type and learning of multiple feature importance rules through computed action of an XCS.

Performance improvements of 3.5%, 13% and 5.7% were observed in terms of area under PR, F-measure and classification accuracy curves by the XCSCA method. The results of a two-sided t-test confirmed the statistical significance of the results.

Analysis of the grouping schemes revealed that the images placed in the same group by the XCSCA method exhibit lower spread in terms of mean absolute error (MAE) performance metric, as compared with the MAE spread of the grouping obtained by the IGA method. Grouping similarly performing images resulted in more generalised feature importance rules for the XCSCA method. Also the IGA method was not able to detect individual feature level variations due to comparison of whole feature vectors. On the other hand, the XCSCA method is capable of detecting individual feature level changes due to its relative encoding scheme.

Analysis of the categorical results revealed that the IGA method explicitly performs better than the XCS based method on images with a simple background, conversely the XCS based method specifically shows better performance than the IGA method on scenes with difficult backgrounds.



The XCSCA method is more suitable for learning multiple combination schemes as compared to the IGA method as it can provide more generalisation with less training and comparable test times. However, if the image features employed require a search for appropriate normalisation and integration schemes, the IGA method may be preferred.

### 10.2.3 Dynamic Feature Selection (DFS)

Novel cues were introduced to dynamically select appropriate maps for combination and neglect unwanted maps in chapter 8. The goal of dynamically selecting only appropriate maps for combination was successfully achieved by measuring feature quality based on new quality measuring cues as proposed by the DFS method.

The DFS method advanced the current state-of-the-art in saliency aggregation. The multi-level saliency fusion and the saliency model aggregation experiments suggested that the proposed method was able to identify redundant unwanted maps falsely included in the fusion schemes of methods, thereby improving upon benchmark salient object detection methods. Specifically, the introduction of DFS in the fusion scheme of DRFI resulted in performance improvements of 2.8% in terms of maximum F-measure, while it improved the performance of PW method by 4.3% in terms of maximum F-measure.

The DFS based salient object detection method introduced in chapter 8 advanced the current state in salient object detection by improving upon several methods. This was achieved by identifying and discarding noisy features in difficult images with cluttered background and multiple distractor salient objects.

In addition to benefiting benchmark methods, the DFS method benefited the FGSopt and IGA methods proposed in this thesis as discussed in chapter 9. The DFS based improvements in feature combination come at the cost of additional computational time spent by the DFS in judging

features. However, various trade-offs between accuracy and time in terms of DFS variants are presented (see chapter 9). The DFS variants do not include computationally expensive cues for achieving computational efficiency at the cost of decreased salient object detection accuracy. Therefore, a variant of the DFS method must be chosen according to the desired application task and specific accuracy and timing constraints.

## **10.3 Future Work**

### **10.3.1 Matting Based Figure-ground Segregation**

The FGS method employed various methods to speed up the optimisation process and explored image down-sampling in order to achieve a practical computation time. For real time operation of the FGS method, new methods are required to substantially accelerate the process of matting based feature computation. Also, as the computation time is greatly influenced by the window size of the affinity function, new affinity functions must be explored to support window scaling.

The FGS method employed an unsupervised selection process of the matting components to compute the matting based feature. Despite robust performance of the unsupervised selection process, in rare cases, corrupted information is included in the final FGS response. Considering the high object coverage of the matting components, a supervised method for matting component selection is worthy of further investigation. Methods to reduce the computational overhead of a supervised technique must be explored in parallel.

### **10.3.2 Learning Multiple Combination Rules**

The IGA method employed the euclidean distance metric to perform the k-nearest neighbour search for semi-autonomous image grouping. The results revealed that the simple nature of feature grouping did not result

in grouping highly similar features in terms of their performance as compared with the XCSCA method. Therefore investigation of new and effective grouping schemes for the IGA method are suggested in future.

A computed action based XCS method was investigated to learn only the feature importance weights. In contrast an XCS with code fragmented action has the added capability to learn rules based on simple arithmetic functions. As normalisation and integration schemes are composed of basic mathematical functions, it will be interesting to learn various normalisation and integration schemes with an XCS with code fragmented action.

Due to high computational overhead of training evolutionary methods such as multiple Genetic Algorithms and XCS, smaller benchmark datasets are employed to evaluate these methods. It is an important future direction to apply these learning methods to next generation large scale image datasets such as ImageNet [119]<sup>1</sup> for salient object detection and related applications. A major issue to resolve would be to reduce the computational overhead at the training stage in accordance with the increased training set size.

### 10.3.3 Dynamic Feature Selection

The introduced DFS method has the ability to provide a robust measure of feature quality based on the devised cues. The devised cues also attempt to cover an approximately complete cue space. However the system is not capable of detecting corner cases such as images containing no salient object(s) and image having complex occlusions. Therefore it will be interesting to explore further cues to extend the method to cover such instances.

---

<sup>1</sup>ImageNet is a large-scale benchmark dataset win image recognition with around 1 million images and 1000 object categories. It is used in an international large scale visual recognition challenge to access state-of-the-art methods since 2010.



# Bibliography

- [1] ABEEL, T., VAN DE PEER, Y., AND SAEYS, Y. Toward a gold standard for promoter prediction evaluation. *Bioinformatics* 25, 12 (2009), i313–i320.
- [2] ACHANTA, R., ESTRADA, F., WILS, P., AND SÜSSTRUNK, S. Salient Region Detection and Segmentation. In *Computer Vision Systems*. Springer, 2008, pp. 66–75.
- [3] ACHANTA, R., HEMAMI, S., ESTRADA, F., AND SUSSTRUNK, S. Frequency-tuned Salient Region Detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 1597–1604.
- [4] ACHANTA, R., SHAJI, A., SMITH, K., LUCCHI, A., FUA, P., AND SÜSSTRUNK, S. Slic superpixels. Tech. rep., École Polytechnique Fédéral de Laussanne (EPFL), Tech. Rep, 2010.
- [5] ACHANTA, R., AND SUSSTRUNK, S. Saliency Detection Using Maximum Symmetric Surround. In *Image Processing (ICIP), 2010 17th IEEE International Conf on* (2010), pp. 2653–2656.
- [6] AHLUWALIA, M., AND BULL, L. A Genetic Programming Based Classifier System. In *Proceedings of the Genetic and Evolutionary Computation Conference* (1999), pp. 11–18.
- [7] AL-SAHAF, H., ZHANG, M., JOHNSTON, M., AND VERMA, B. Image descriptor: A genetic programming approach to multiclass tex-

- ture classification. In *Proceedings of 2015 IEEE Congress on Evolutionary Computation* (2015), pp. 2460–2467.
- [8] ALEXE, B., DESELAERS, T., AND FERRARI, V. What is an Object? In *IEEE Conference on Computer Vision and Pattern Recognition* (2010), pp. 73–80.
- [9] ALPERT, S., GALUN, M., BASRI, R., AND BRANDT, A. Image Segmentation by Probabilistic Bottom-up Aggregation and Cue Integration. In *IEEE Conference on Computer Vision and Pattern Recognition* (2007), pp. 1–8.
- [10] ARBELAEZ, P., PONT-TUSET, J., BARRON, J., MARQUES, F., AND MALIK, J. Multiscale combinatorial grouping. In *IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 328–335.
- [11] BERNADÓ-MANSILLA, E., AND GARRELL-GUIU, J. M. Accuracy-Based Learning Classifier Systems: Models, Analysis and Applications to Classification Tasks. *Evolutionary Computation* 11, 3 (2003), 209–238.
- [12] BERTSEKAS, D. P. *Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series)*, 1 ed. Athena Scientific, 1996.
- [13] BORJI, A. Boosting bottom-up and top-down visual features for saliency estimation. In *IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 438–445.
- [14] BORJI, A., PARKS, D., AND ITTI, L. Complementary effects of gaze direction and early saliency in guiding fixations during free viewing. *Journal of Vision* 14, 13 (2014), 1–32.
- [15] BORJI, A., SIHITE, D. N., AND ITTI, L. Salient object detection: A benchmark. In *European Conference on Computer Vision* (2012), pp. 414–429.

- [16] BOYD, S., AND VANDENBERGHE, L. *Convex Optimization*. Cambridge University Press, 2004.
- [17] BRAITHWAITE, R., AND BHANU, B. Hierarchical gabor filters for object detection in infrared images. In *IEEE Conference on Computer Vision and Pattern Recognition* (1994), pp. 628–631.
- [18] BUTZ, M. V. XCSJava 1.0: An Implementation of the XCS Classifier System in Java. Tech. Rep. 2000027, Illinois Genetic Algorithms Laboratory, 2000.
- [19] BUTZ, M. V., LANZI, P. L., AND WILSON, S. W. Function Approximation With XCS: Hyperellipsoidal Conditions, Recursive Least Squares, and Compaction. *IEEE Transactions on Evolutionary Computation* 12, 3 (2008), 355–376.
- [20] BUTZ, M. V., PELIKAN, M., LLORÀ, X., AND GOLDBERG, D. E. Automated Global Structure Extraction for Effective Local Building Block Processing in XCS. *Evolutionary Computation* 14, 3 (2006), 345–380.
- [21] BUTZ, M. V., AND WILSON, S. W. An Algorithmic Description of XCS. *Soft Computing* 6, 3-4 (2002), 144–153.
- [22] CARREIRA, J., AND SMINCHISESCU, C. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 7 (2012), 1312–1328.
- [23] CERF, M., HAREL, J., EINHÄUSER, W., AND KOCH, C. Predicting human gaze using low-level saliency combined with face detection. In *Advances in neural information processing systems* (2008), pp. 241–248.

- [24] CHANG, C.-C., AND LIN, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27.
- [25] CHANG, K.-Y., LIU, T.-L., CHEN, H.-T., AND LAI, S.-H. Fusing generic objectness and visual saliency for salient object detection. In *IEEE International Conference on Computer Vision* (2011), pp. 914–921.
- [26] CHENG, M.-M., WARRELL, J., LIN, W.-Y., ZHENG, S., VINEET, V., AND CROOK, N. Efficient salient region detection with soft image abstraction. In *IEEE International Conference on Computer Vision* (2013), pp. 1529–1536.
- [27] CHENG, M.-M., WARRELL, J., LIN, W.-Y., ZHENG, S., VINEET, V., AND CROOK, N. Efficient salient region detection with soft image abstraction. In *IEEE International Conference on Computer Vision* (2013), pp. 1529–1536.
- [28] CHENG, M.-M., ZHANG, G.-X., MITRA, N., HUANG, X., AND HU, S.-M. Global contrast based salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2011), pp. 409–416.
- [29] CHENG, M.-M., ZHANG, G.-X., MITRA, N. J., HUANG, X., AND HU, S.-M. Global Contrast Based Salient Region Detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2011), pp. 409–416.
- [30] COMANICIU, D., AND MEER, P. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 5 (2002), 603–619.
- [31] DAM, H. H., ABBASS, H. A., LOKAN, C., AND YAO, X. Neural-Based Learning Classifier Systems. *IEEE Transactions on Knowledge and Data Engineering* 20, 1 (2008), 26–39.



- [32] DEEP, K., SINGH, K. P., KANSAL, M., AND MOHAN, C. A real coded genetic algorithm for solving integer and mixed integer optimization problems. *Applied Mathematics and Computation* 212, 2 (2009), 505–518.
- [33] DEEP, K., SINGH, K. P., KANSAL, M. L., AND MOHAN, C. A real coded genetic algorithm for solving integer and mixed integer optimization problems. *Applied Mathematics and Computation* 212, 2 (2009), 505–518.
- [34] DEEP, K., AND THAKUR, M. A new crossover operator for real coded genetic algorithms. *Applied Mathematics and Computation* 188, 1 (2007), 895–911.
- [35] DEEP, K., AND THAKUR, M. A new mutation operator for real coded genetic algorithms. *Applied Mathematics and Computation* 193, 1 (2007), 211–230.
- [36] DOLLAR, P., AND ZITNICK, C. Structured forests for fast edge detection. In *IEEE International Conference on Computer Vision* (2013), pp. 1841–1848.
- [37] DONOSER, M., AND BISCHOF, H. Diffusion processes for retrieval revisited. In *IEEE Conference on Computer Vision and Pattern Recognition* (June 2013), pp. 1320–1327.
- [38] ESHELMAN, L. J., AND SCHAFFER, J. D. Real-coded genetic algorithms and interval-schemata. In *Foundation of Genetic Algorithms* 2 (1993), Morgan Kaufmann., pp. 187–202.
- [39] EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C., WINN, J., AND ZISSERMAN, A. The pascal visual object classes (voc) challenge. *IJCV* 88, 2 (2010), 303–338.

- [40] FILIPE, S., ITTI, L., AND ALEXANDRE, L. A. Bik-bus: Biologically motivated 3d keypoint based on bottom-up saliency. *IEEE Transactions on Image Processing* 24, 1 (2015), 163–175.
- [41] FORGY, E. W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* 21 (1965), 768–769.
- [42] FREY, B. J., AND DUECK, D. Clustering by passing messages between data points. *Science* 315, 5814 (2007), 972–976.
- [43] FRINTROP, S. *VOCUS: a visual attention system for object detection and goal-directed search*, Ph.D. dissertation. PhD thesis, Rheinische Friedrich-Wilhelms-Universitt, 2005.
- [44] FRINTROP, S., KLODT, M., AND ROME, E. A real-time visual attention system using integral images. In *In Proceedings of the 5th International Conference on Computer Vision Systems (ICVS (2007))*.
- [45] FU, H., CAO, X., AND TU, Z. Cluster-based co-saliency detection. *IEEE Trans. Image Process.* 22, 10 (2013), 3766–3778.
- [46] GOFERMAN, S., ZELNIK-MANOR, L., AND TAL, A. Context-aware saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2010), pp. 2376–2383.
- [47] GOLDBERG, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st ed. Addison-Wesley Longman Publishing Co., Inc., 1989.
- [48] GOLDBERG, D. E. *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Kluwer Academic Publishers, 2002.
- [49] GOPALAKRISHNAN, V., HU, Y., AND RAJAN, D. Salient region detection by modelling distributions of colour and orientation. *IEEE Trans. Multimedia* 11, 5 (2009), 892–905.

- [50] HAREL, J., KOCH, C., AND PERONA, P. Graph-based visual saliency. In *NIPS* (2006), pp. 545–552.
- [51] HOLLAND, J. H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. MIT Press, 1992.
- [52] HOLTZMAN-GAZIT, M., ZELNIK-MANOR, L., AND YAVNEH, I. Salient Edges: A Multi Scale Approach. In *European Conference on Computer Vision, Workshop on Vision for Cognitive Tasks* (2010).
- [53] HONG, Y., AND ZHU, W. Spatial co-training for semi-supervised image classification. *Pattern Recognition Letters* 63 (2015), 59–65.
- [54] HOU, X., AND ZHANG, L. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition* (2007), pp. 1–8.
- [55] HOU, X., AND ZHANG, L. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition* (2007), pp. 1–8.
- [56] HU, Y., XIE, X., MA, W.-Y., CHIA, L.-T., AND RAJAN, D. Salient region detection using weighted feature maps based on the human visual attention model. In *Advances in Multimedia Information Processing-PCM 2004*. Springer Berlin Heidelberg, 2005, pp. 993–1000.
- [57] HUBEL, D. H., AND WIESEL, T. N. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of physiology* 160 (1962), 106–154.
- [58] HURVICH, L. M., AND JAMESON, D. An opponent-process theory of colour vision. *Psychological review* 64, 6 (1957), 384–404.

- [59] IOANNIDES, C., BARRETT, G., AND EDER, K. Improving XCS Performance on Overlapping Binary Problems. In *Proceedings of the IEEE Congress on Evolutionary Computation* (2011), pp. 1420–1427.
- [60] IQBAL, M., BROWNE, W., AND ZHANG, M. Reusing building blocks of extracted knowledge to solve complex, large-scale boolean problems. *Evolutionary Computation, IEEE Transactions on* 18, 4 (2014), 465–480.
- [61] IQBAL, M., BROWNE, W. N., AND ZHANG, M. XCSR with Computed Continuous Action. In *Proceedings of the Australasian Joint Conference on Artificial Intelligence* (2012), pp. 350–361.
- [62] IQBAL, M., BROWNE, W. N., AND ZHANG, M. Evolving Optimum Populations with XCS Classifier Systems. *Soft Computing* 17, 3 (2013), 503–518.
- [63] IQBAL, M., BROWNE, W. N., AND ZHANG, M. Extending Learning Classifier System with Cyclic Graphs for Scalability on Complex, Large-Scale Boolean Problems. In *Proceedings of the Genetic and Evolutionary Computation Conference* (2013), pp. 1045–1052.
- [64] IQBAL, M., BROWNE, W. N., AND ZHANG, M. Learning Complex, Overlapping and Niche Imbalance Boolean Problems Using XCS-Based Classifier Systems. *Evolutionary Intelligence* 6, 2 (2013), 73–91.
- [65] ITTI, L., AND KOCH, C. Feature Combination Strategies for Saliency-Based Visual Attention Systems. *Journal of Electronic Imaging* 10, 1 (2001), 161–169.
- [66] ITTI, L., KOCH, C., AND NIEBUR, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1254–1259.

- [67] JIANG, B., ZHANG, L., LU, H., YANG, C., AND YANG, M.-H. Saliency detection via absorbing markov chain. In *IEEE International Conference on Computer Vision* (2013), pp. 1665–1672.
- [68] JIANG, H., WANG, J., YUAN, Z., WU, Y., ZHENG, N., AND LI, S. Salient object detection: A discriminative regional feature integration approach. In *IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 2083–2090.
- [69] JIANG, P., LING, H., YU, J., AND PENG, J. Salient region detection by ufo: Uniqueness, focusness and objectness. In *IEEE International Conference on Computer Vision* (2013), pp. 1976–1983.
- [70] JONES, J. P., AND PALMER, L. A. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology* 58, 6 (1987), 1233–1258.
- [71] JUDD, T. *Understanding and predicting where people look in images*. PhD thesis, Massachusetts Institute of Technology, 2011.
- [72] JUDD, T., EHINGER, K., DURAND, F., AND TORRALBA, A. Learning to predict where humans look. In *IEEE International Conference on Computer Vision* (2009), pp. 2106–2113.
- [73] KENDALL, M. G. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.
- [74] KENNEDY, J., AND EBERHART, R. C. *Swarm Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.
- [75] KIENZLE, W., WICHMANN, F. A., FRANZ, M. O., AND SCHÖLKOPF, B. A nonparametric approach to bottom-up visual saliency. In *NIPS* (2007), pp. 689–696.

- [76] KIM, J., HAN, D., TAI, Y.-W., AND KIM, J. Salient region detection via high-dimensional colour transform. In *IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 883–890.
- [77] KLEIN, D., AND FRINTROP, S. Center-surround Divergence of Feature Statistics for Salient Object Detection. In *IEEE International Conference on Computer Vision* (2011), pp. 2214–2219.
- [78] KOCH, C., AND ULLMAN, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology* 4, 4 (1985), 219–227.
- [79] KOCH, C., AND ULLMAN, S. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*. Springer Netherlands, 1987, pp. 115–141.
- [80] KOWALIW, T., BANZHAF, W., AND DOURSAT, R. Networks of transform-based evolvable features for object recognition. In *Genetic and Evolutionary Computation Conference, GECCO '13, Amsterdam, The Netherlands, July 6-10, 2013* (2013), pp. 1077–1084.
- [81] KOZA, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. The MIT Press, 1992.
- [82] KRIG, S. *Computer Vision Metrics: Survey, Taxonomy, and Analysis*, 1st ed. Apress, Berkely, CA, USA, 2014.
- [83] KUKENYS, I., BROWNE, W. N., AND ZHANG, M. Transparent, On-line Image Pattern Classification Using a Learning Classifier System. In *Applications of Evolutionary Computation*, vol. 6624. Springer Berlin Heidelberg, 2011, pp. 183–193.
- [84] LANZI, P. L., AND LOIACONO, D. Classifier Systems That Compute Action Mappings. In *Proceedings of the Genetic and Evolutionary Computation Conference* (2007), pp. 1822–1829.

- [85] LANZI, P. L., LOIACONO, D., WILSON, S. W., AND GOLDBERG, D. E. Generalization in the XCSF Classifier System: Analysis, Improvement, and Extension. *Evolutionary Computation* 15, 2 (2007), 133–168.
- [86] LEUNG, T., AND MALIK, J. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV* 43, 1 (2001), 29–44.
- [87] LEVIN, A., LISCHINSKI, D., AND WEISS, Y. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 2 (2008), 228–242.
- [88] LEVIN, A., RAV-ACHA, A., AND LISCHINSKI, D. Spectral matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 10 (2008), 1699–1712.
- [89] LI, B., PETERSON, M. R., AND FREEMAN, R. D. Oblique effect: A neural basis in the visual cortex. *Journal of Neurophysiology* 90, 1 (2003), 204–217.
- [90] LI, J., M.D., L., AN, X., XU, X., AND HE, H. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 99 (2012), 1–16.
- [91] LI, X., LU, H., ZHANG, L., RUAN, X., AND YANG, M.-H. Saliency detection via dense and sparse reconstruction. In *IEEE International Conference on Computer Vision* (2013), pp. 2976–2983.
- [92] LI, Y. *Hypergraph Modeling for Saliency Detection and Beyond*. PhD thesis, The University of Adelaide, 2013.
- [93] LIU, R., CAO, J., LIN, Z., AND SHAN, S. Adaptive partial differential equation learning for visual saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 3866–3873.

- [94] LIU, T., YUAN, Z., SUN, J., WANG, J., ZHENG, N., TANG, X., AND SHUM, H.-Y. Learning to Detect a Salient Object. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 2 (2011), 353–367.
- [95] LIU, Y., MU, C., KOU, W., AND LIU, J. Modified particle swarm optimization-based multilevel thresholding for image segmentation. *Soft Comput.* 19, 5 (2015).
- [96] LOIACONO, D., MARELLI, A., AND LANZI, P. L. Support Vector Machines for Computing Action Mappings in Learning Classifier Systems. In *Proceedings of the IEEE Congress on Evolutionary Computation* (2007), pp. 2141–2148.
- [97] LU, S., MAHADEVAN, V., AND VASCONCELOS, N. Learning optimal seeds for diffusion-based salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 2790–2797.
- [98] MA, Y.-F., AND ZHANG, H.-J. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the Eleventh ACM International Conference on Multimedia* (2003), pp. 374–381.
- [99] MAI, L., AND LIU, F. Comparing salient object detection results without ground truth. In *European Conference on Computer Vision* (2014), vol. 8691, pp. 76–91.
- [100] MAI, L., NIU, Y., AND LIU, F. Saliency aggregation: A data-driven approach. In *IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 1131–1138.
- [101] MARGOLIN, R., TAL, A., AND ZELNIK-MANOR, L. What makes a patch distinct? In *IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 1139–1146.
- [102] MARTIN, D., FOWLKES, C., TAL, D., AND MALIK, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics.



- In *IEEE International Conference on Computer Vision* (2001), vol. 2, pp. 416–423.
- [103] MOVAHEDI, V., AND ELDER, J. Design and perceptual validation of performance measures for salient object segmentation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2010), pp. 49–56.
- [104] NAQVI, S., BROWNE, W., AND HOLLITT, C. Optimizing visual attention models for predicting human fixations using genetic algorithms. In *IEEE Congress on Evolutionary Computation* (2013), pp. 1302–1309.
- [105] NAQVI, S. S., BROWNE, W. N., AND HOLLITT, C. Combining Object-Based Local and Global Feature Statistics for Salient Object Search. In *IVCNZ* (2013), pp. 394–399.
- [106] NAQVI, S. S., BROWNE, W. N., AND HOLLITT, C. Optimizing Visual Attention Models for Predicting Human Fixations Using Genetic Algorithms. In *Proceedings of IEEE Congress on Evolutionary Computation* (2013), pp. 1302–1309.
- [107] NAVALPAKKAM, V., AND ITTI, L. An integrated model of top-down and bottom-up attention for optimal object detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2006), pp. 2049–2056.
- [108] NEIL, D. B. B., AND TSOTSOS, J. K. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision* 9, 3 (2009), 1–24.
- [109] PALMER, S. E. *Vision science : photons to phenomenology*. MIT Press, 1999.

- [110] PARKS, D., BORJI, A., AND ITTI, L. Augmented saliency model using automatic 3d head pose detection and learned gaze following in natural scenes. *Vision Research* (2015), 1–12.
- [111] PELE, O., AND WERMAN, M. Fast and Robust Earth Mover's Distances. In *Proceedings of International Conference on Computer Vision* (2009), pp. 460–467.
- [112] PENG, B., AND LIMING, Z. Visual saliency: a biologically plausible contourlet-like frequency domain approach. *Cognitive Neurodynamics* 4 (2010), 189–198.
- [113] PERAZZI, F., KRAHENBUHL, P., PRITCH, Y., AND HORNUNG, A. Saliency filters: Contrast based filtering for salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 733–740.
- [114] PETERS, R. J., IYER, A., ITTI, L., AND KOCH, C. Components of bottom-up gaze allocation in natural images. *Vision Research* 45, 8 (2005), 2397–2416.
- [115] POSNER, M. I., AND COHEN, Y. Components of Visual Orienting. *Attention and Performance X* 32 (1984), 531–556.
- [116] RATHA, N. K., JAIN, A. K., AND LAKSHMANAN, S. Object detection in the presence of clutter using gabor filters. In *SPIE's 1994 International Symposium on Optics, Imaging, and Instrumentation* (1994), pp. 612–623.
- [117] RICHE, N., MANCAS, M., GOSSELIN, B., AND DUTOIT, T. RARE: a New Bottom-Up Saliency Mode. In *Proceedings of the International Conference on Image Processing (IEEE ICIP)* (2012), pp. 641–644.
- [118] RUBNER, Y., TOMASI, C., AND GUIBAS, L. J. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision* 40, 2 (2000), 99–121.

- [119] RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATY, A., KHOSLA, A., BERNSTEIN, M., BERG, A. C., AND FEI-FEI, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* (2015), 1–42.
- [120] SASTRY, K., GOLDBERG, D., AND KENDALL, G. Genetic algorithms. In *Search Methodologies*. Springer US, 2005, pp. 97–125.
- [121] SEO, H. J., AND P., M. Nonparametric bottom-up saliency detection by self-resemblance. In *Computer Vision and Pattern Recognition Workshops. IEEE Computer Society Conference on* (2009), pp. 45–52.
- [122] SEO, H. J., AND P., M. Static and space-time visual saliency detection by self-resemblance. *Journal of vision* 9, 12 (2009), 1–27.
- [123] SHEN, X., AND WU, Y. A unified approach to salient object detection via low rank matrix recovery. In *IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 853–860.
- [124] SPEARMAN, C. "General Intelligence," objectively determined and measured. *The American Journal of Psychology* 15, 2 (1904), 201–292.
- [125] THAKOOR, K., N. MANTE, SIAGIAN, C., WEILAND, J. D., ITTI, L., AND MEDIONI, G. A system for assisting the visually impaired in localization and grasp of desired objects. In *European Conference on Computer Vision, Workshop on Assistive Computer Vision and Robotics, Zurich, Switzerland* (2015), pp. 643–657.
- [126] TILKE, J., FRÉDO, D., AND ANTONIO, T. A benchmark of computational models of saliency to predict human fixations. Tech. Rep. MIT-CSAIL-TR-2012-001, MIT Computer Science and Artificial Intelligence Lab (CSAIL), January 2012.

- [127] TONG, N., LU, H., ZHANG, Y., AND RUAN, X. Salient object detection via global and local cues. *Pattern Recognition* 48, 10 (2014), 3258–3267.
- [128] TORRALBA, A., OLIVA, A., CASTELHANO, M., AND HENDERSON, J. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review* 113, 4 (2006), 766–786.
- [129] TRAN, T. H., SANZA, C., DUTHEN, Y., AND NGUYEN, D. T. XCSF with Computed Continuous Action. In *Proceedings of the Genetic and Evolutionary Computation Conference* (2007), pp. 1861–1869.
- [130] TREISMAN, A. M., AND GELADE, G. A feature-integration theory of attention. *Cognitive Psychology* 12, 1 (1980), 97–136.
- [131] TSOTSOS, J. K. A 'complexity level' analysis of immediate vision. *International Journal of Computer Vision* 1, 4 (1988), 303–320.
- [132] VALOIS, R. L. D., ALBRECHT, D. G., AND THORELL, L. G. Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research* 22, 5 (1982), 545–559.
- [133] VIG, E., DORR, M., AND COX, D. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 2798–2805.
- [134] WALTHER, D., AND KOCH, C. Modeling attention to salient proto-objects. *Neural Networks* 19, 9 (2006), 1395–1407.
- [135] WANG, S., AND SISKIND, J. Image segmentation with ratio cut. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 6 (2003), 675–690.

- [136] WEI, Y., WEN, F., ZHU, W., AND SUN, J. Geodesic saliency using background priors. In *European Conference on Computer Vision* (2012), pp. 29–42.
- [137] WILSON, S. W. Classifier Fitness Based on Accuracy. *Evolutionary Computation* 3, 2 (1995), 149–175.
- [138] WILSON, S. W. Mining Oblique Data with XCS. In *Proceedings of the Genetic and Evolutionary Computation Conference (Companion)* (2000), pp. 158–174.
- [139] WILSON, S. W. Classifiers that Approximate Functions. *Natural Computing* 1 (2002), 211–233.
- [140] WILSON, S. W. Three Architectures for Continuous Action. In *Learning Classifier Systems*. Springer Berlin Heidelberg, 2007, pp. 239–257.
- [141] WISCHNEWSKI, M., STEIL, J. J., KEHRER, L., AND SCHNEIDER, W. X. Integrating inhomogeneous processing and proto-object formation in a computational model of visual attention. *Cognitive Systems Monograph* 6 (2009), 93–102.
- [142] WITTEN, I. H., AND FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., 2005.
- [143] WOLFE, J. M., CAVE, K. R., AND FRANZEL, S. L. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance* 15, 3 (1989), 419.
- [144] XIE, Y., LU, H., AND YANG, M.-H. Bayesian saliency via low and mid level cues. *IEEE Trans. Image Process.* 22, 5 (2013), 1689–1698.

- [145] XUE, B., ZHANG, M., BROWNE, W., AND YAO, X. A survey on evolutionary computation approaches to feature selection. *Evolutionary Computation, IEEE Transactions on PP*, 99 (2015), 1–1.
- [146] XUE, B., ZHANG, M., AND BROWNE, W. N. A comprehensive comparison on evolutionary feature selection approaches to classification. *International Journal of Computational Intelligence and Applications* 14, 02 (2015), 1550008.
- [147] YAN, Q., XU, L., SHI, J., AND JIA, J. Hierarchical saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 1155–1162.
- [148] YANG, C., ZHANG, L., LU, H., RUAN, X., AND YANG, M.-H. Saliency detection via graph-based manifold ranking. In *IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 3166–3173.
- [149] YANG, C., ZHANG, L., LU, H., RUAN, X., AND YANG, M.-H. Saliency detection via graph-based manifold ranking. In *IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 3166–3173.
- [150] ZHAI, Y., AND SHAH, M. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th Annual ACM International Conference on Multimedia* (2006), pp. 815–824.
- [151] ZHANG, J., AND SCLAROFF, S. Saliency detection: A boolean map approach. In *IEEE International Conference on Computer Vision* (2013), pp. 153–160.
- [152] ZHANG, L., TONG, M. H., MARKS, T. K., SHAN, H., AND COTRELL, G. W. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision* 8 (2008), 32–32.

- [153] ZHAO, Q., AND KOCH, C. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision* 11, 3 (2011), 1–15.
- [154] ZHAO, Q., AND KOCH, C. Learning visual saliency by combining feature maps in a nonlinear manner using adaboost. *Journal of Vision* 12, 6 (2012).
- [155] ZHOU, D., WESTON, J., GRETTON, A., BOUSQUET, O., AND SCHÖLKOPF, B. Ranking on data manifolds. In *NIPS* (2003), pp. 169–176.
- [156] ZHOU, Z.-H. Learning with unlabeled data and its application to image retrieval. In *PRICAI 2006: Trends in Artificial Intelligence*. Springer Berlin Heidelberg, 2006, pp. 5–10.
- [157] ZHU, L., KLEIN, D., FRINTROP, S., CAO, Z., AND CREMERS, A. A multisize superpixel approach for salient object detection based on multivariate normal distribution estimation. *IEEE Trans. Image Process.* 23, 12 (2014), 5094–5107.
- [158] ZHU, W., LIANG, S., WEI, Y., AND SUN, J. Saliency optimization from robust background detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 2814–2821.