

Understanding Visual Saliency and Perception of Digital Media

by

Shafin Rahman

A thesis submitted to
The Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements
of the degree of

Master of Science

Department of Computer Science
The University of Manitoba
Winnipeg, Manitoba, Canada
August 2015

© Copyright 2015 by Shafin Rahman

Thesis advisor

Neil D. B. Bruce

Author

Shafin Rahman

Understanding Visual Saliency and Perception of Digital Media

Abstract

Eye movements are directed by both high-level considerations such as contextual guidance or experience and also by observed visual patterns. A bright light or high contrast image structure will tend to draw an observer's attention, independent of task. The extent to which a visual pattern attracts attention in this manner is often described according to a quantitative measure of *visual saliency*. Although many computational models of visual saliency have been published in last two decades, current literature is mostly limited to the maximization of performance in predicting saliency. Apart from this common challenge, we address some new aspects of image understanding relating to saliency. For example, we explore the relationship between natural image statistics and human visual attention, measure the extent of center bias produced by saliency algorithms and develop ways to remove such bias, relate different perceptual task like free viewing, object search, saliency search and explicit judgment, and predict different image level subjective ratings. In this thesis, we present solutions to aforementioned problems for a better basic understanding of visual saliency while also producing some novel avenues of analysis that inform on perception of digital media.

Contents

Abstract	ii
Table of Contents	iv
List of Figures	v
List of Tables	vii
Acknowledgments	viii
Dedication	ix
1 Introduction	1
1.1 General introduction	1
1.2 Motivation and contribution	2
1.3 Thesis organization	6
2 Related Work	7
2.1 Saliency Models	7
2.1.1 Bottom-up approach	8
2.1.2 Contextual Guidance	10
2.1.3 Spectral analysis approach	11
2.1.4 Learning-based approaches	12
2.1.5 Region or object search approach	13
2.2 Natural image statistics	14
2.3 Center Bias	16
2.4 Evaluation of saliency predictions	17
2.5 Explicit Judgment	18
2.6 Inter observer congruency	20
2.7 Subjective image rating	21
3 Saliency Prediction	24
3.1 Natural image saliency	24
3.1.1 Improvement of human-like visual saliency prediction	25
3.1.2 Saliency benchmarking on natural image	30
3.1.3 Discussion	32

3.2	Center bias	33
3.2.1	Extent of center bias in saliency maps	34
3.2.2	Removal of center bias	35
	Rank order based center bias removal	37
	Whitening based center bias removal	40
3.2.3	Discussion	40
3.3	Explicit judgment	43
3.3.1	Benchmarking of explicit judgment	43
3.3.2	Approximation of explicit judgments	46
3.3.3	Comparison of tasks in terms of IOC	47
3.3.4	Relation of explicit judgment to IOC in fixation tasks	49
3.3.5	Explicit judgment prediction model	51
3.3.6	Explicit judgment performance prediction	57
3.3.7	Details of post-processing blur for ensemble prediction	59
3.3.8	Saliency and Segmentation	59
3.3.9	Discussion	61
4	Digital Media Rating	63
4.1	Inter-observer congruency prediction	63
4.1.1	IOC computation	64
4.1.2	Experimental Methods and Results	65
	Single diagnostic features	67
	More complex features	70
4.1.3	Discussion	77
4.2	Other subjective ratings	79
4.2.1	Prediction model	81
4.2.2	t-stochastic neighbor embedding (tSNE) of image features	81
4.2.3	Experimental results	82
4.2.4	Discussion	86
5	Summary, Limitations and Future Work	89
5.1	Summary	89
5.2	Limitations	90
5.3	Future Work	91
	Bibliography	100

List of Figures

3.1	Various stages of the SOC processing cascade for a naturalistic image	27
3.2	Spatial bias map of algorithms on Koehler <i>et al.</i> [29] dataset (a) Torralba [49], (b) HouCVPR [16], (c) HouNIPS [17], (d) AWS [13], (e) Itti-CIO2 [19], (f) ImageSignatureLab [15], (g) ImageSignatureRGB [15], (h) SDSR [46], (i) AIM [8], (j) GBVS [14], (k) Yan [56] and (l) SOC [40].	35
3.3	Rank order based center bias removal on Toronto dataset	36
3.4	Rank order based center bias removal on Judd dataset	37
3.5	Rank order based center bias removal on Koehler <i>et al.</i> [29] dataset	38
3.6	Rank order based center bias removal on Koehler <i>et al.</i> [29] dataset (Explicit Judgment)	39
3.7	Whitening based center bias removal on Toronto dataset	41
3.8	Whitening based center bias removal on Judd dataset	42
3.9	Whitening based center bias removal on Koehler <i>et al.</i> [29] dataset	43
3.10	Whitening based center bias removal on Koehler <i>et al.</i> [29] dataset (Explicit Judgment)	44
3.11	Scatter plot showing the correlation among the IOC of free view, obj. search and sal. view tasks	49
3.12	Scatter plot showing the correlation between Standard AUC of explicit judgment approximation vs. IOC scores of (a) free view (b) object search (c) saliency view fixation data	51
3.13	Performance for explicit judgment prediction. (Best viewed in digital format)	56
3.14	Explicit judgment prediction using different feature sets.	56
3.15	Sample output column (a) input image (b) ex. judgment locations (c) predicted ex. judgment map (d) simulated ex. judgment locations (e) free view fixation locations (f) predicted fixation map	57
3.16	Precision Recall Curve for one and two object case using both explicit judgment and fixation map	60

4.1	Images are sorted based on their actual IOC score from (a) Bruce (b) Judd (c) Memorability datasets. Top rows corresponding to each dataset depict those images having low IOC and bottom rows depict images having high IOC. Specific properties common among higher or lower IOC scores at a semantic level are discussed in detail throughout this section.	66
4.2	Predicted IOC vs. actual IOC for each combination of the three datasets, and HoG, GIST and HoPS features as predictors. The red line corresponds to the least squares linear regression fit to the data. .	75
4.3	Images among the greatest 15 underestimates of actual IOC scores (based on residuals from the regression model): (a) Common to both HoPS and Deep learning features (b) Within the bottom 15 for HoPS only. (c) Within the bottom 15 for Deep learning only.	78
4.4	2D Embedding of deepNet feature by tSNE. DeepNet feature captures mostly the high level features. Because of this dominant objects of similar nature (like human, car) are grouped together in the embedding	82
4.5	2D Embedding of image Gist feature by tSNE. It is quite noticeable that global color directs the embedding. As Gist is applied on raw image, color becomes the distinguishing factor	83
4.6	2D Embedding of saliency Gist feature by tSNE. Images having similar looking saliency map are embedded in close locations	83
4.7	2D Embedding of image HoG feature by tSNE. Images having similar local edge structure are in close location	84
4.8	2D Embedding of saliency HoG feature by tSNE. Images having similar looking edges in saliency map are embedded in close locations	84
4.9	2D Embedding of image histogram feature by tSNE. This embedding is dominated by both color and local structure	85
4.10	2D Embedding of saliency HoPS feature by tSNE. Images having similar looking saliency map are embedded in close locations	85

List of Tables

3.1	Performance of SOC model at different stage on DOVES dataset . . .	28
3.2	Comparison with some of the best algorithms in the literature	30
3.3	Benchmarking performance for the Koehler <i>et al.</i> [29] dataset	45
3.4	Prediction of explicit judgment using Gaussian blurred free view, ob- ject search and saliency viewing fixation map respectively.	52
3.5	For a fixed NTree = 15, the effect of different Gaussian blur level in the explicit judgment prediction model	59
3.6	F-measure of segmentation performance	60
4.1	Correlation between actual IOC scores for single diagnostic features determined to have significant correlation with IOC.	70
4.2	Correlation between actual IOC score and JPEG image size subject to varying degrees of JPEG quality	70
4.3	Correlation between actual IOC scores and predicted IOC scores using different <i>holistic</i> feature vectors for prediction.	71
4.4	Prediction performance as Pearson Correlation of subjective rating of different types of features	86
4.5	Testing with combination of features for rating type Brand Experience	87
4.6	Testing with combination of features for rating type Relevance	87
4.7	Testing with combination of features for rating type Valence	87
4.8	Testing with combination of features for rating type Originality . . .	88

Acknowledgments

I would like to thank my supervisor Dr. Neil Bruce. His careful guidance has enabled me to reach this point. His thoughtful insight to solving problems has continuously helped me to find every solution presented in this thesis. I am thankful to the computer vision lab and to my fellow lab members for creating and maintaining a wonderful lab environment. I would like to express great gratitude to my parents and other family members for supporting and encouraging me throughout the duration of my degree. A special thanks goes to my beloved wife for sacrificing my presence while I was living away from home for study. Finally, I acknowledge the financial support I have received from the Department of Computer Science and University of Manitoba Graduate Fellowship.

This thesis is dedicated to my parents.

Chapter 1

Introduction

1.1 General introduction

The human visual system is well adapted to finding areas of interest in a scene. In most of the cases, unexpected phenomena like a flash of light, colorful items, objects in motion or a location that is dissimilar to its surroundings attract our interest. This distinct perceptual behavior is called image salience. Simulating this attentive behavior using a computational model is an open challenge of research [23]. Models of this type produce a saliency map of an input image, with high values corresponding to more salient pixel locations, and lower values to less salient pixel locations. This representation can be used as input to other algorithms for scene analysis, or for visualization in highlighting the salient regions. Models attempt to maximize the match between the saliency maps produced, and observed fixations in eye tracking data from user studies. This saliency map can be used as part of a larger set of components in systems for image navigation, object detection, segmentation,

recognition etc.

1.2 Motivation and contribution

Established algorithms for characterizing image saliency have informed on different aspects of visual attention, and have gradually produced superior performance in predicting gaze data. In most cases, algorithms have been tested with datasets containing many salient objects. However, for a natural image dataset where relatively few salient objects or regions are present, the algorithms have not been tested exhaustively. In such a case, the attentive behaviour of humans will vary significantly from one person to another because of scarcity of salient patterns. In this thesis, we analyse model behavior in considering natural images to better understand behaviour in more naturalistic settings. Furthermore, for natural stimuli, some prior models show the capability to predict the neural response corresponding to visual processing [12; 27] as measured via brain imaging. As the response is related to visual attention (or rather stimulus contrast), we attempt to provide human like saliency predictions with a model inspired by these brain imaging efforts.

Human gaze behaviour is largely dominated by a common tendency called center bias [47]. People have a tendency to fixate preferentially towards the middle of the scene. We can easily find the effect of this behaviour in eye tracking fixation data for almost any dataset. The extent of center bias inherent in this kind of data has been addressed in detail in the literature [4]. Different saliency algorithms treat center bias according to a variety of methods within the saliency detection process. For example, many algorithms apply Gaussian blurring as a post processing step which

implicitly models the effect of center bias [34]. In other research efforts, it is shown that algorithms are very sensitive to this type of blurring [4; 15]. As the amount of center bias inherent in the output of different algorithms varies, it is important to measure the extent of center bias to correctly interpret benchmarking results. To solve this problem, a modified type of ROC analysis has been proposed that considers area under the ROC curve (AUC) subject to random selection of negative samples; this has been proposed to remove the effect of center bias during evaluation [47]. However, there is also spatial bias in the output of different algorithms in the saliency map itself, and this has not been addressed in the literature. In this thesis, we propose a way to measure the extent of center bias produced by different saliency algorithms. Moreover, the role of center bias in benchmarking and evaluation of models has not been considered in sufficient detail, and we propose alternative ways to treat center bias in the saliency maps to ensure a fair comparison among saliency algorithms.

While fixation data captures gaze behavior through eye tracking, this presents one specific measure of content of relevance to an observer. There are alternatives that are arguably subject to a different degree of cognitive control. One such measurement proposed in prior work is deemed explicit judgment. For explicit judgment, the most salient location in an image is selected by a human observer via manual selection using the mouse, and is subject to a slower time-course that presumably invokes a different set of cognitive processes. Koehler *et al.* [29] showed that saliency maps better approximate explicit judgments than gaze points in fixation data. In this thesis, we present results that improve on performance of algorithmic prediction of locations likely to be selected via explicit judgment. We have built a new model for prediction

of explicit judgment based on an ensemble prediction using different saliency maps and comparing with fixation data. This includes methods for approximating the type of data that results from measuring explicit judgment without a user study in simulating this type of data from extant eye tracking data. This research expands on our understanding of visual perception of digital media, and also for alternative types of analysis that include reduces the impact of biases associated with oculomotor control, or center bias.

In eye tracking experiments, each observer will tend to examine different locations in an image. As mentioned before, the collection of those locations are characterized by discrete fixations. In fixation data, there is typically correlation among the individual locations fixated by different observers, however the extent of this correlation will vary from one image to another. In the saliency literature, the extent of the agreement across observers is measured and quantified according to an inter-observer congruency (IOC) score [31; 23]. This plays a role primarily in benchmarking studies in measuring the capacity for the raw human fixation data to predict behavior of other humans in the experiments, and to provide a bound on how well an algorithm may perform in the best case [4] since this score is bounded by the inherent variability in the data. Aside from this traditional use in benchmarking, IOC may arguably provide a good measure of characteristics of an image, including the level of interest an image may generate. A high IOC means that observers are inclined to look at similar locations in the image. Such agreement indicates that observers have an implicit tendency to examine the image in a common fashion that may be due to salient elements of the image. There is therefore value in knowing the degree of IOC

associated with an image, and therefore to having the capacity to infer this statistic in the absence of existing eye tracking data [32]. IOC prediction can be an useful tool as an intermediate level feature for analytics targeting perception of digital media. In this thesis, we have proposed a computational model for predicting IOC when eye tracking data is unavailable, and we suggest some simple (entropy, visual clutter and JPEG image size) and complex (in from of HoG, Gist, histogram of predictive salience and deep learning) diagnostic features that vary in their relationship to bottom-up and top-down processing.

In the same spirit as measuring IOC, some other image level subjective ratings of human observers can also be predicted in a similar fashion. In existing literature, predictions of subjective qualities of media has been considered for video and/or text level [28] representations. There are a handful of existing studies that consider image level ratings including image popularity [28] and memorability [18]. In this thesis, we have explored some other subjective ratings like brand experience, purchase intention, interest, understanding, valance, relevance and originality for images of magazine advertisements. We have explored different bottom-up and top-down features to predict those subjective ratings. Predicting these subjective ratings has high value in perception of digital media, especially in the advertising industry. There is also insight the derives from this analysis that is applicable to alternative forms of visual media, and prediction of subjective dimensions of media in general.

1.3 Thesis organization

The remainder of this thesis is organized as follows: In Chapter 2, background and related work is introduced. This includes a description of different types of saliency detection models and their evaluation criteria, and analysis of the impact of center bias in evaluation. Chapter 2 also introduces a variety of background relevant to topics covered in this thesis including statistics of natural images, details of measuring explicit judgment, and measuring inter-observer congruency and subjective image qualities. Chapter 3 demonstrates the performance of saliency algorithms on naturalistic images, motivates a saliency model inspired by human vision, and considers various factors related to benchmarking including estimation of the extent of center bias and removal of such bias, prediction of explicit judgments and analysis of different perceptual tasks. Chapter 4 addresses prediction of different types of image level perceptual and subjective ratings including inter-observer congruency, brand experience, purchase intention, interest, understanding, valance, relevance and originality. Finally, Chapter 5 concludes the dissertation describing limitations and directions for future research.

Chapter 2

Related Work

2.1 Saliency Models

The salient locations in a scene depend on the composition of the scene including areas of contrast, figure ground relationships and prior knowledge. Different saliency algorithms produce differences in output dependent on the specific computational stages that define them. The modeling assumptions for any model, also vary in their motivation and on the nature of low-level pattern contrast that defines salient content. Thus, it is very important to review proposed models in the literature that are successful in revealing locations of interest. Based on differences in solution strategy, visual saliency methods can be grouped into different categories. Categories considered in this chapter include bottom-up, contextual guidance, spectral analysis, learning based and region search based approaches. The nature of these categories is somewhat evident from their names, but greater clarity on these definitions is presented in the remainder of this chapter. Additional review papers covering this

subject can be found in [4], [23] and [7].

2.1.1 Bottom-up approach

Bottom-up approaches model visual attention as a free viewing task without any particular goal. For example, a series of images is presented to an observer without providing any particular instruction about what to search for in the images. In such conditions, approach to model overt visual attention is bottom-up and driven by characteristics of the image. This type of strategy is assumed most models of image saliency. It reveals image understanding related to color, shape, orientation, or motion irrespective of specific objects. One of the early attempts to model saliency by Itti et al. [19] is based on “feature integration theory”. They considered color, intensity and orientation feature maps from different scales as Gaussian pyramids. They calculated the center surround difference by subtracting the pixel values of the bottom layer from upsampled higher layers. The ranges of center surround effect are different for color, shape or orientation and for this they normalized contrast channels to ensure a similar range of effect. After that, they combined each of the color, intensity and orientation maps by summing the normalized center surround maps across different scales of the Gaussian pyramid. The final saliency map is computed by calculating the average of three combined maps for color, intensity and orientation. The strengths of this approach are effectively predicting the bottom-up attention of early visual processes of primates and making it possible to implement massively parallel computation.

Bruce and Tsotsos [8] proposed a theory of attention based on information maxi-

mization (AIM). This approach performs an offline processing stage where the system calculates some independent coefficients to represent local patterns in an image by applying principal component analysis and independent component analysis (ICA) on random image patches from a large database. After that, to process an input image, the system extracts independent features from every patch using the coefficients of ICA. Then, the feature density is estimated by making the normalized histogram of features. For every pixel of the input image, the system finds the joint likelihood of each coefficient from the histogram to calculate the self-information. Finally, the authors model the self-information to compute a saliency map corresponding to the input image.

Another popular saliency algorithm is Graph-Based Visual Saliency (GBVS) [14]. This algorithm works in three steps: Extraction, activation and normalization. The extraction step is similar to traditional feature maps based on color, intensity, orientation and biologically motivated filtering. In the activation step, each pixel is assumed to be a node of a graph and connected to its surrounding neighboring pixels by an edge. The weight of this edge is the product of pixel value dissimilarity and closeness of the neighboring pixel spatially. In this way, the authors define a Markov chain by normalizing the weights of outward edges of a node to 1. In the normalization step, with the activation values the authors propose a similar Markov chain to normalize the activation of a node by the activation of the neighbor nodes and their spatial distance. The output of this normalization step is assumed to correspond to the saliency value of the pixel. This approach is a different way of considering non-linearity and normalization required for saliency detection, which also helps parallel

implementation.

In another work, Garcia-Diaz *et al.* [13] introduced a decorrelation based method for visual saliency. They calculated local orientation energy and color features and applied principal component analysis on those multiscale features, dividing maps corresponding to principal components by their associated eigenvalues. Then, they combined all features maps by simple addition. To obtain the final result they also consider saliency normalization and Gaussian smoothing. Yan *et al.* [56] proposed a saliency model using matrix decomposition of image features. They used both spatial and feature domain statistics to learn a dictionary. This dictionary is used to represent the image from patch-wise features. Finally, image saliency is produced by rank sparsity decomposition of feature norm. Hou and Zhang [17] showed an image saliency model by continuous sampling of features. They used the principle of predictive coding strategy to explain saliency. By incremental coding length, they calculated energy of image pixels and predicted the fixation points.

2.1.2 Contextual Guidance

In the real world, our attention is not always a free viewing task. In most of the cases, contextual information motivates our search behaviour in paying attention to scenes. For example, if an image contains a person, the observer may intend to look at person's face. Here, the semantic relevance of the face takes precedence, even if there is no specific task involved. The contextual guidance approach takes advantage of prior knowledge of the scene to predict fixation points. Therefore, in addition to pattern driven salience, such approaches also consider task specific visual search.

Models including context typically involve object specific understanding of scene, or a representation of holistic scene structure. One of the early contributions of this kind is proposed by Torralba *et al.* [49; 51] which is based on a Bayesian framework. By calculating local and global features from image, this model demonstrates the impact of contextual guidance in image saliency. By combining bottom up saliency and contextual modulation, the model produces a modulated saliency map based on holistic scene structure. This model can also find the regions of a scene likely to contain a certain salient object. Zhan *et al.* [57] also demonstrated a Bayesian framework for contextual guidance to predict the locations of fixation points. They considered bottom up saliency using self-information, and top-down activation as term based on log-likelihoods of task relevant patterns combining these to determine saliency. In this case, they did not consider any global statistics. Kanan *et al.* [25] combined the two previously described approaches and the proposed object appearance based saliency model. If the target object is previously known, this model can predict target specific image saliency.

2.1.3 Spectral analysis approach

Models based on spectral analysis compute visual saliency in the Fourier domain by finding the log spectrum of the image for difference scales. Image understanding in the frequency domain may also help to locate areas of interest. Hou and Zhang [16] published the first model of this kind. This model did not consider any image features, categories or any prior information for contextual guidance, only calculating the log spectrum and a spectral residual based on a smoothed version of the log

spectrum. Wang and Li [55] extended this model by incorporating an automatic channel selection and a decision reversal module. One of the major drawbacks of such a strategy is that the computation has little fidelity to biological correlates of saliency.

Another widely used model of this kind is image signature based saliency [15]. The method assumes that an image signal is a mixture of foreground and background. The goal is to reconstruct the image isolating the support of foreground. This approach applies the Discrete Cosine Transform (DCT) on different channels of the input image separately. Then, it takes the *sign* of the DCT output which is called the image signature. After that, the salient image is constructed by taking the mean of the inverse DCT of the image signature corresponding to different colour channels. As a post processing, the method applies Gaussian blur on the constructed image because it is known that all different saliency algorithms assume optimal performance in the predictions they form with blurring. The authors experimented using both RGB and Lab channels of images and found image signature based on Lab performs better than RGB.

2.1.4 Learning-based approaches

Learning-based approaches models visual saliency detection as a machine learning problem. This strategy typically models image saliency as binary classification task. A trained system can classify each pixel of the input image as either a salient or non-salient pixel. The eyetracking data of different images from many human observer are used to train the system. Seo and Milanfar [45] suggested a learning-based saliency

model based on a self-resemblance measure. This model calculated local regression features from each pixel to capture local structure of the data. By modeling likelihoods through kernel density estimation, this model determines the saliency map. In another model, Zhao and Koch [58] propose to first find color, intensity, orientation and face conspicuity feature map for each pixel of the input image and subsequently learned weights associated with those features. They defined an objective function which is minimized by the least square technique to classify salient pixels.

2.1.5 Region or object search approach

All the strategies described above find the salient locations from the scene and compare those locations with eye tracking data of human observers. But those strategies can not tell whether the location belongs to any object or any background pixel of the image. To address this problem, object search approaches find the boundaries of salient object pixels instead of only locating fixation points. This strategy takes advantage of image saliency for object detection. Achanta *et al.* [1] proposed a salient region detection method by using frequency domain computation. This approach found the boundary of the object which can preserve the maximum frequency content of the entire image. Another significant contribution with a similar strategy is established by Liu *et al.* [35], where they modelled the boundary search of an object as a machine learning problem. This approach learned a conditional random field based on the image characteristics like contrast of multi-scale decomposition, center-surround histogram and spatial distribution of color. According to the training, this approach located the salient object within the scene.

2.2 Natural image statistics

Humans are able to determine if an image is natural or artificial very rapidly because all natural scenes have common features. The statistics that define the structure of natural images are often referred to using the term natural image statistics. Examining the statistics of natural images shows that the structure of such images is not random and in the entire domain of all possible images, natural images only reside within a very small subspace. Neural computation within the visual system of primates exploits this structure in representing visual patterns. A consequence of this is that natural image statistics have characteristics that are encoded efficiently by neurons and may be represented according to simple mathematical functions motivated by this statistical relationship. Many computer vision research efforts attempt to discover statistical regularities or structure in image patterns. One of the early attempts of this kind was done by Field [11] who showed consistent statistical properties of natural images and their relationship to the responses of visual cortical cells. More specifically, units in early visual cortex are well represented by the output of Gabor functions. These function are localized in both time and frequency with the ability to represent any time varying signal. Each of the functions can be obtained by the multiplication of Gaussian and sinusoid function. The mammalian visual system encodes a natural scene through neurons and create a sparse representation of the scene by maximizing information and reducing redundancy of environmental data. Field's results suggest that a sparse, non-redundant image coding scheme may be produced through Gabor functions that are analogous to early neural processing for vision. This is a famous linear model for neural coding. Later, Schwartz et al. [44]

showed that although the linear model has a strong link between natural signal and neural behaviour, such a model cannot imitate the non-linear behaviour of sensory neurons. They argued that as light comes from different surfaces of the natural scene, the sensory neurons combine light in a non-linear fashion. They applied two linear filters having same the frequency tuning but different orientation tuning on a natural image and observed that for a fixed spatial location, the output of the first filter can predict the output of the second filter. This means that both of the filter responses are mutually dependent. But, in our visual system, any two neurons behave independently. To solve this issue, Schwartz et al. [44] proposed a non-linear model to reduce the statistical dependency of the standard linear model. These simplistic models are successful in predicting blood oxygenation level dependent (BOLD) responses of neurons in viewing natural images and checkerboard patterns, however responses get saturated for low contrast energy. An improvement to this approach is proposed by Kay et al. [27] wherein divisive normalization (DN) and second order contrast (SOC) follow the computation of total spectral energy to achieve better prediction of BOLD responses. Therefore, energy filters coupled with DN and SOC provide a characterization for predicting the response of neurons to naturalistic stimuli.

The neural response is greatly altered by the top-down contextual guidance or task associated with the natural stimuli. Torralba *et al.* [50] examined how top-down contextual information affects Gabor like filter output. He showed that the image scale as well as observers interaction with the scene affects observed image statistics. Different image categories also imply different predictable statistical regularities within a scene. Image statistics of both natural and man-made environments are

compared in examining the average power spectrum of the image Fourier transform within different categories. By varying the distance between the observer and the scene, statistics are altered such that image categories with man-made objects vary only in horizontal and vertical contours of the spectrum. On the other hand, natural image categories have significant variation spread across the spectral signature of the image regardless of orientation.

2.3 Center Bias

Human gaze behaviour is largely dominated by a common tendency called center bias. People have a tendency to fixate preferentially towards the middle of the scene. We can easily find the effect of this behaviour in eye tracking fixation data for almost any dataset. Tatler [47] discussed in detail cause and effect of center bias in human eye tracking data. He suggested reasons for center bias which include motor biases in the oculomotor system, bias in the distribution of image features, viewing strategy of subjects, set-up of the eye tracking experiment and other situational factors. The motor bias of our visual system is the phenomenon of favoring smaller amplitude saccades over large amplitude saccades. If the presentation time of the stimuli is very small then the motor bias will become more dominant. Moreover, while collecting eye tracking data, a pre-trial fixation marker is normally placed at the center of the screen. The joint effect of motor bias and center marker may bias the observer towards viewing the center. According to Tatler, another important factor is the feature distribution within the image. He found that fixation data strongly correlates with low level image features like brightness. If the image is taken in such a way that

it exhibits a feature distribution skewed towards center, the associated distribution of fixations will surely contain center bias. Tatler argued that fixation behavior is also motivated by task dependence which can be viewed as top down information considered while searching a scene. In case of task dependence involved, the fixated locations more closely follow the regions where target objects are present.

It is important to measure the extent of center bias in fixation data because it helps to avoid the problem of data bias influencing assessment of performance of saliency algorithms. In a recent study, Borji *et al.* [4] proposed a center-bias ratio method to quantify the degree of bias. This method imagines circles of different radii centered in the image and calculates the ratio of number of fixation points inside the circle to the total number of fixations. In this way, for all circle sizes this method finds a vector of ratio values. Vector values then provide a diagnostic for the degree of central bias. If the fixation data contains strong center bias then the first few values of the vector will have high quantity. Using a suitable threshold Borji *et al.* determine the degree of center bias.

2.4 Evaluation of saliency predictions

In most cases, saliency algorithms are evaluated by calculating area under the curve (AUC) corresponding to ROC analysis (in plotting true positive rate vs. false positive rate). This is the standard AUC method of evaluating algorithms [24]. However, this way of evaluating algorithms is not fair because it does not consider the center bias present in the fixation data. To solve this problem, Tatler *et al.* [48] proposed a shuffled AUC (sAUC) method for evaluation. In this approach, for any

given image, positive samples are the fixation data of all users for that specific image and negative samples are the fixation data of all users across all other images in the dataset. By choosing the positive and negative samples based on the overall fixation data, this evaluation method implicitly considers the nature of fixation distribution and evaluates an algorithm without the effect of center bias. Again, many bottom up saliency methods do not model the effect of center bias in the algorithms. For this reason, those algorithms apply a Gaussian blur stage at the end of the process. This blurring implicitly considers the effect of center bias and helps to increase the evaluation measure score. Different algorithms may benefit most from different degrees of blurring which can also make the comparison of algorithms unfair [34].

2.5 Explicit Judgment

Fixation data is recorded by an eye tracker and is largely dominated by the nature of stimulus patterns, bottom-up, top-down and other external factors sometimes subject to a rapid time-course and with some observations outside of cognitive control. Because of the mixed effect of all these factors, fixation data may be seen as contaminated in the sense that it exhibits spatial bias, poor localization of objects, and other noise factors. For this reason, evaluating a saliency algorithm based on fixation data may present some limitations. To address this issue, one reliable alternative for saliency evaluation involves explicitly selecting locations/regions of interest manually instead of measuring gaze locations. The selected location/region is called the explicit judgment of the observer corresponding to the most salient location [5; 29]. Borji *et al.*[5] investigates explicit judgment at region or object level. They performed a user

study to let the participants choose any of the two objects presented in an image of the SED dataset [2] by clicking inside the selected object or drawing a polygon around the object boundary. They found that both fixation and bottom-up saliency models can predict selections that follow this type of judgment of salient content. They also reported that fixation prediction models can predict salient judgments better than standard salient object detection models.

Again, in another recent study, Koehler *et al.* [29] pointed out the shortcomings of fixation data to predict saliency and for the first time, published a rich dataset containing explicit judgments. This study considers the explicit judgment task at the pixel level. Each of the 100 participants were asked to select one pixel per image which corresponds to the location that stands out most within the scene. They reported the performance of three well-known fixation based saliency algorithms in their ability to approximate the explicit judgment locations. They found that instead of predicting this explicit judgment, all well-known saliency models attempt to predict the gaze points in fixation data. Apart from the explicit judgment task, this work also involved three other tasks i.e. free viewing, saliency and a cued object search task. The free viewing task represents searching behavior without any particular goal, which is the scenario traditionally considered in most datasets in the literature. In the saliency searching task, observers were asked to decide whether the right or left part of the image is more salient than the other. Finally, in the object searching task, observers were instructed to find some pre-defined objects.

2.6 Inter observer congruency

Looking at an image, different people may find different parts of the image to be of interest. For this reason, image saliency algorithms attempt to produce a saliency map that is aligned with the locations of common interest to viewers. However, after recording many viewers' eye tracking data, it is possible that there is little agreement in terms of common locations of interest. In such a case, the performance of saliency algorithms is necessarily low as the salient locations vary significantly from person to person. To address this issue, prior research attempted to find a maximum upper bound on performance for predicting gaze points across a set of images [4; 31; 23]. This bound is established by using the eye tracking data from many observers to predict gaze points of an additional observer. This upper bound of performance is termed the inter observer congruency (IOC) score. This score is normally used to provide a reference frame for performance of algorithms that predict likelihoods of locations of gazed-upon coordinates in a scene.

The possibility of predicting IOC score by a computational model without the use of data from a human study is an interesting idea. Le Meur *et al.* [32] addressed this idea for the first time in the literature. They discussed several underlying reasons for inter-observer variability like cultural differences, and prior experience of observers. They also noted that the more time a scene is viewed for, the more variability within fixation locations is observed. They proposed the first computational model to predict the IOC score. Considering the factors that influence IOC, they suggested 6 different features to predict IOC i.e. the presence of faces, color harmony, depth of field, entropy based scene complexity, color mean shift segmentation, and amount of contours.

Using these features from an image, they trained a cluster-weighted model which is a generalization of Gaussian mixture. For quantitative evaluation, they calculated the correlation between predicted and actual IOC values. Their reported correlation is not significant within a low confidence interval. They mentioned several facts regarding the limitations such as a small number of images in the dataset, the influence of viewing duration and low performance of the face detector. Again, they argued that IOC within the free viewing data is less predictable than task specific fixation data like object search. The only similar work in this line of investigation is by Rosenholtz *et al.* [42] but instead of predicting IOC, they measured visual clutter which is the level of ease to add a target that will draw attention within an existing image. Le Meur *et al.* [32] confirmed that visual clutter is highly correlated with IOC.

2.7 Subjective image rating

Multimedia content can be rated according to a wide variety of different subjective criteria. Many studies have already been undertaken on video rating prediction with the help of visual content and text mining. However, only a few of these efforts are based on image level (as opposed to pixel level) subjective rating where only visual features are used in prediction. There is evident value for applications to have a sense of expected ratings without knowing the general comments on the visual content, or conducting a user study. Some of the most recent research on such ideas are based on image popularity [28] and memorability [18].

Now-a-days people are uploading millions of photos to the internet, especially via social networking websites. Among the huge number of photos, a few of these are

viewed by a very large number of people, whereas many of the others are completely ignored by people. It is evident that some photos must have some characteristics which make those distinct within the huge number of available photos, or there exist characteristics that define the continuum between these two extremes. Khosla *et al.* [28] attempted to find those characteristics to predict the popularity of images. They experimented on 2.3 million Flickr photos and considered popularity based on number of views by people. They found that both image content and social context contribute to the popularity of an image. To extract image context they used Gist, texture statistics, color patches, gradients, and deep learning features as low level image content and use object features representative of different categories as high level image content. Moreover, they also considered some social cues in the form of mean views, photo count, contacts, groups, member duration etc. of any given user, number of tags, title length, description length of any photo. Using all the features from both image content and social cues, they trained a support vector regression (SVR) model to predict number of views (i.e. popularity) associated with the test photo. They reported that the prediction results are highly correlated with the true popularity value.

Like popularity, another subjective (or non-semantic) criterion is image memorability. Isola *et al.* [18] consider memorability of images, and claim that this is one of the more stable properties of an image. Some images become memorable because of known people or special moments of our lifetime, but there are also some other images which become memorable without the presence of these elements. Isola *et al.* [18] addressed memorability associated with this second factor. To find a memorability

value for each image, the authors designed a memory game for the participants of a user study. In the study, participants are asked to find the images from an image sequence those were previously shown to them. Then, the memorability score for any specific image is calculated as the percentage of correct detections of that image in the whole experiment. Finally, they experimented with different types of features like color, object and global features to predict the memorability score using SVR model.

Chapter 3

Saliency Prediction

3.1 Natural image saliency

Visual saliency models have largely been tested on gaze data from this type of experiment, but images appearing in most data sets tend to have significant context, objects or salient regions that may imply the involvement of higher-level cognitive routines. For this reason, it is important to understand the behavior of models within more naturalistic settings wherein there are weaker prior on context, or objects present. To this end, in this section, we evaluate a number of algorithms that have performed well in recent benchmarking studies, subjected to natural images. For such naturalistic stimuli, it is also natural to consider computational saliency from the perspective of visual information processing in human brain. To this end, we also test a recent simplistic biologically motivated model that is supported by its agreement with brain imaging data [12]. A simplistic bio-inspired model of early neural activation in the visual cortex is based on log-Gabor filters with activation corresponding to the

sum of spectral energy represented across filters corresponding to a particular location [12]. This model is motivated by the complex cell energy model wherein energy corresponds to paired oriented Gabor-like filters in quadrature. This simplistic model is successful in predicting blood oxygenation level dependent (BOLD) responses in viewing natural images and checkerboard patterns, however responses get saturated for low contrast energy. An improvement to this approach is proposed by Kay et al. [27] wherein divisive normalization (DN) and second order contrast (SOC) follow the computation of total spectral energy to achieve better prediction of BOLD responses. Therefore, it is of interest to examine the extent to which energy filters coupled with DN and SOC provides a characterization of fixated locations for naturalistic stimuli. This model has therefore also been implemented and included in our comparison.

3.1.1 Improvement of human-like visual saliency prediction

The primary visual cortex (V1) of the human visual system is important in basic feature extraction, and thus plays an important role in visual attention. Gabor-like filters may be used to model the response of cells in V1 and also to naturalistic stimuli. This provides a basis for examining computational determination of visual saliency from the perspective of the visual information processing in the human brain. A great effort has been performed for predicting the responses of V1 neuron [12]. These responses can be blood oxygenation level dependent (BOLD). Many recent researches have proved that V1 energy of BOLD responses can be modeled by Gabor like filter. Actually, LogGabor filter works well for statistical analysis of natural images. For this reason, keeping agreement with visual system, the first step of attention modeling

is to apply LogGabor filters on input grayscale image and sum-up all filter responses to find a saliency map. The following equation is the frequency responses of LogGabor filter:

$$F_{v1}(x, y) = \exp \frac{-\log(x'/y')^2}{2 \log(\sigma_x/x')^2} \cdot \exp \frac{-y'^2}{2\sigma_y^2} \quad (3.1)$$

where, $x' = x \cos(\theta) + y \sin(\theta)$, $y' = -x \sin(\theta) + y \cos(\theta)$, θ is the orientation of the filter, x' is the center frequency and σ_x and σ_y are the bandwidth of the filter along the x and y direction respectively. Using different scale and orientation a bank of filter can be produced. After that, summing up all filter responses we can get the basic saliency map.

We can improved the described model using some steps from a recent studies presented in [27]. The prediction of BOLD responses using Gabor like filter can not model the fact that responses saturate at low contrasts. To solve this problem, Kay *et al.* [27] proposed divisive normalization (DN). This step is able to capture several nonlinear response properties of V1 neurons. Computationally, this step can be performed by the following equation:

$$D_n = \frac{G_{sc,or}^r}{s^r + \left(\frac{\sum_{sc,or} G_{sc,or}}{S \times O} \right)^r} \quad (3.2)$$

where, $G_{sc,or}$ is the LogGabor filter responses for s th scale and o th orientation; S and O are the total number of scale and orientation respectively; s and r controls the strength of the normalization.

This model is still incomplete because it overestimate the actual BOLD responses. We have to apply variance of contrast energy to get the actual responses. This is called

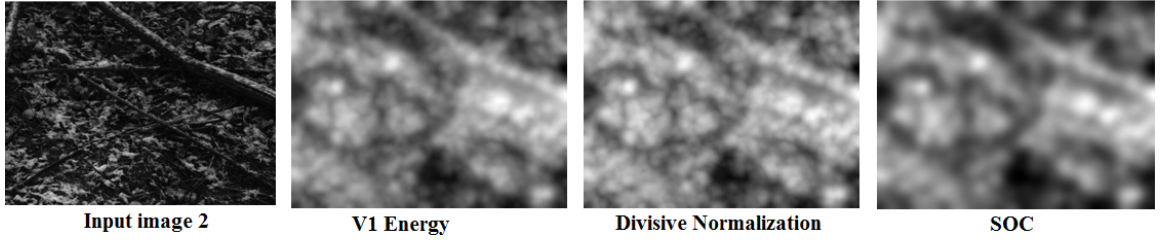


Figure 3.1: Various stages of the SOC processing cascade for a naturalistic image second order contrast step. To incorporate variance like nonlinearity we apply the following equation:

$$SOC = \sum_i w_i (D_n - c \sum_j w_j D_j) \quad (3.3)$$

where, w is 2D Gaussian weights, D_j is the Gaussian filtered image of D_i and c controls the strength of non-linearity of second order contrast. SOC contains the final saliency map.

We have implemented this model and compared the performance with 11 other established algorithms on several image datasets. Again, both standard ROC [23] and shuffled ROC [4; 57] analysis are reported in evaluation in Table 3.1. We have named this model the SOC model. It shows that DN and SOC step after basic log-Gabor filtering (i.e. V1 energy) improves the performance compared to the absence of these additional stages of computation. Various stages of the SOC processing cascade for a naturalistic image are showed in Figure 3.1.

Implementation: In the context of the SOC model spectral energy is characterized by log-Gabor filters in quadrature, and subject to four parameters (minimum wavelength, a multiplier that defines the centre frequency for any given scale, the σ

Table 3.1: Performance of SOC model at different stage on DOVES dataset

Evaluation	V1 Energy	Divisive Normalization	SOC Model
Standard AUC	0.572318	0.572636	0.573822
Shuffled AUC	0.551890	0.552003	0.555988

of the Gaussian envelope containing the log-Gabor filter and total number of scales). The total number of orientation was fixed to 6 as behavior of the algorithm was found to be relatively insensitive to this parameter. Parameters for energy filtering were selected from a broad sample of parameter combinations to produce a best fit to the characterization of fixation locations on the DOVES dataset. In particular, minimum wavelength was chosen from 2, 3, 6, 12 pixels, the multiplier associated with peak frequency of bandpass filters from one scale to the next 1.4, 1.7, 2.2, the sigma on the Gaussian envelope (normalized by frequency) from .75, .65, .55 and number of scales from 3, 4. The best parameter set was found to be 12, 1.7, 0.65, 4 for the four respective parameters. It is important to note that this provides even spectral coverage over a broad band of radial frequencies. Subsequently, parameters for divisive normalization (as defined in the SOC model) involved optimizing r and s in choosing the best r from .7, .9, 1, 1.2 and s from .5, 4, 8, 16. The best results were obtained in choosing $r = .7$ and $s = 8$. Finally second order contrast (SOC) parameters were optimized with the 2 parameters σ and c optimized selecting the best c among 0.5, 0.7, 0.9 and σ among 2, 4, 8, 16. Optimal results were obtained in choosing $\sigma = 2$, $c = 0.5$. As an additional test of fidelity (and to avoid the possibility that parameters are overfit to a specific dataset, we have also compared the results against the best performing algorithms in prior benchmarks on the Toronto dataset

with the same parameters optimized for the DOVES dataset. A comparison of performance of algorithms that fare well in the benchmark of Borji et al. in addition to the SOC model is presented in Table 3.2. We have use two different datasets for this comparison, specifically the DOVES and Toronto dataset, with the latter serving as an additional test of the simplistic SOC model. It is evident that the performance for the DOVES dataset is significantly lower than that of the Toronto dataset. This most likely reflects the observation that the DOVES dataset is relatively devoid of very salient objects. It is also the case that the performance of the relatively simplistic SOC model is within a similar range to some of the better performing saliency algorithms, despite its simplicity. It is important to also note that the SOC model operates only on a grayscale version of the Toronto dataset, and is therefore at a disadvantage in information provided as input but nevertheless performs in a range that is proximal to the state-of-the-art.

It is important to note that the large differences in the Standard AUC scores may be attributed primarily to the extent to which there is central bias in the algorithms evaluated. Based on the shuffled AUC score, we have computed the inter-observer average auROC for the DOVES dataset to be 0.5526 ± 0.04518 with a maximum of 0.6418, and for the Toronto dataset to be 0.7372 ± 0.0918 , with a maximum of 0.8768. This suggests that the saliency algorithms tested are within the range of the inter-observer prediction scores for the naturalistic data, but fall short of the inter-observer predictions for the Toronto data. Inter-observer predictions were computed on a leave-one-out basis for all observer/image combinations, with fixations from remaining observers convolved with a Gaussian to produce a saliency map. The

optimal σ for the Gaussian convolution was determined based on the average auROC across all individual/image combinations to provide with the best σ on average used to compute IO scores. This corresponded to a σ of 3% of the image width. Detail description of inter-observer score is presented in Chapter 4.

3.1.2 Saliency benchmarking on natural image

In last few decades, many saliency algorithms have been published claiming very good performance across different datasets. In most cases, those datasets include images with prominent objects, faces, texts, and oddball patterns. However, there is no benchmarking effort in the case where the dataset does not contain any significant number of salient objects. In general, naturalistic images do not contain many highly salient locations. For this reason, we wish to investigate some of the best saliency algorithms and their performance on a naturalistic image database.

Table 3.2: Comparison with some of the best algorithms in the literature

Algorithms [4]	DOVES dataset		Toronto dataset	
	Standard AUC	Shuffled AUC	Standard AUC	Shuffled AUC
Torralba [49]	.582	.556	.825	.673
HouCVPR [16]	.557	.551	.780	.675
HouNIPS [17]	.655	.539	.780	.661
AWS [13]	.534	.508	.537	.537
Itti-CIO2 [19]	.808	.566	.808	.647
ImageSignatureLab [15]	.612	.550	.815	.698
ImageSignatureRGB [15]	.612	.549	.796	.677
SDSR [46]	.543	.553	.814	.707
AIMf [8]	.562	.559	.827	.685
Gbvs [14]	.777	.549	.825	.615
Yanf [56]	.691	.553	.810	.671
SOC (Inspired by [27])	.573	.556	.789	.652

We have selected 12 of the best performing algorithms from the review of Borji *et al.* [4]. These algorithms are applied to the natural images within the DOVES dataset [52]. Performance in predicting fixated regions is examined through area under the curve (AUC) of receiver operating characteristic (ROC) analysis. In AUC based evaluation in this domain, there are two different types of ROC analysis named standard ROC [23] and shuffled ROC [4; 57] that treat spatial bias differently. We have considered both evaluation criteria to compare algorithms. To compare the performance with non-natural image dataset, we also applied all algorithms on Toronto dataset [8]. Results are reported in Table 3.2.

We have found that the established algorithms do not achieve very good performance on naturalistic images (DOVES dataset) compared with images having man-made colorful objects (Toronto dataset). The reason is naturalistic image often contains less salient region that can attract human attention. It motivates people to look at different arbitrary locations in the scene which follows low commonality in viewing. For having low commonality, algorithms cannot able to tune its parameters to predict possible salient locations. Therefore, the inability of prediction does not necessarily mean that algorithms are incapable of predicting saliency. It is only the dataset which does not present sufficient information to let the algorithms perform. We can infer that if the input images contain several context like prominent objects, faces, texts, and oddball patterns, most of the algorithms are most likely to perform well.

3.1.3 Discussion

Given the significant number of potential applications, visual saliency has increasingly become an area of interest in image and vision research. Many different strategies for predicting visual saliency have been proposed, that differ in their composition or rationale, and with a significant focus on improving performance across standard benchmarks. Recent benchmarks considering a large number of algorithms have further provided an understanding of the behavior of different algorithms. Performance evaluation has primarily focused on indoor and outdoor images of urban environments, many of which are composed, and contain salient objects. In this section, we have presented the performance of a number of the better performing algorithms on data derived from naturalistic scenes. In addition, given the strong connection to human vision, we test SOC model for early visual processing in primates tied to spectral energy and normalization. This model is motivated by the complex cell energy model wherein energy corresponds to paired oriented Gabor-like filters in quadrature. This simplistic model is successful in predicting blood oxygenation level dependent (BOLD) responses in viewing natural images and checkerboard patterns, however responses get saturated for low contrast energy. An improvement to this approach is proposed by Kay et al. [27] wherein divisive normalization (DN) and second order contrast (SOC) follow the computation of total spectral energy to achieve better prediction of BOLD responses. Experimental results of SOC model demonstrate significant differences between common datasets, and natural images. In general, natural images have less salient locations and thus, influence algorithms to perform relatively low.

3.2 Center bias

Observers have a tendency to look at the middle of an image or scene, a tendency referred to as center-bias. Tatler [47] discussed in detail the cause and effect of center bias in human eye tracking data. He showed that in free viewing, center bias is present even if salient objects are not within the center of the scene. Reasons for center bias include motor biases in the saccadic system, the distribution of image features, prior bias in the viewing strategy of subjects, the specifics of the experimental setup, and other situational factors [47; 57; 4]. To account for spatial (centre) bias in fixation data, algorithm output is often re-weighted by a centrally located Gaussian profile to better predict fixated regions in the data. However, there are different degrees of systematic spatial bias in the output produced by different algorithms. Since different saliency algorithms exhibit different degrees of center bias, this poses a challenge for producing a fair comparison across algorithms [4]. To address this problem, Zhang *et al.* [57] proposed a metric for evaluation of saliency algorithms called shuffled AUC. In ROC analysis, this evaluation method chooses positive and negative samples in a manner that removes the effect of center bias, by selecting negative samples from a spatial distribution that matches fixations within the entire dataset. However, as a side-effect, this may result in uneven importance of pixel locations in evaluation. Given a much larger number of centrally located samples across the dataset, the importance of saliency output at the center is relatively diminished given that there are likely to be more negative samples within the center region. This diminishes the capacity for an above chance prediction in a signal detection theoretic sense as a function of spatial position. Moreover, for a saliency algorithm having a relative bias

outside of the center allows for the true positive rate to grow more quickly than the false positive rate due to the spatial distribution of negative samples. This issue has not yet been addressed in existing work. In this section, we propose a method to control for bias in both algorithm output and data such that all pixel locations are equally important, and algorithms have no spatial bias.

3.2.1 Extent of center bias in saliency maps

The degree of center bias within data has been recognized as an important factor in performance evaluation. In a recent study, Borji *et al.* [4] proposed a center-bias ratio method to quantify the degree of bias in fixation data. This method assumes circles of different radii centered in the image and calculates the ratio of the number of fixation points inside the circle to the total number of fixations for all possible circle sizes, providing a vector of ratio values that reflects the degree of central bias. Independent of the data, different saliency algorithms may exhibit different degrees of spatial bias due to their inherent computational structure. It is therefore of value to examine spatial bias among saliency algorithms in a manner analogous to the case for fixation data. We visualize this bias as follows: We first whiten individual saliency maps for 12 different algorithms to place the output values of different saliency maps on a common scale (and such that every saliency map has mean of 0 and standard deviation of 1). Subsequently, the mean of all saliency maps produced by a single algorithm is subtracted from the mean across all algorithms. This provides a topological profile of relative spatial bias. This offers a sense of relative center bias in output produced by each algorithm. The overall spatial bias profiles are shown in Figure 3.2. In this

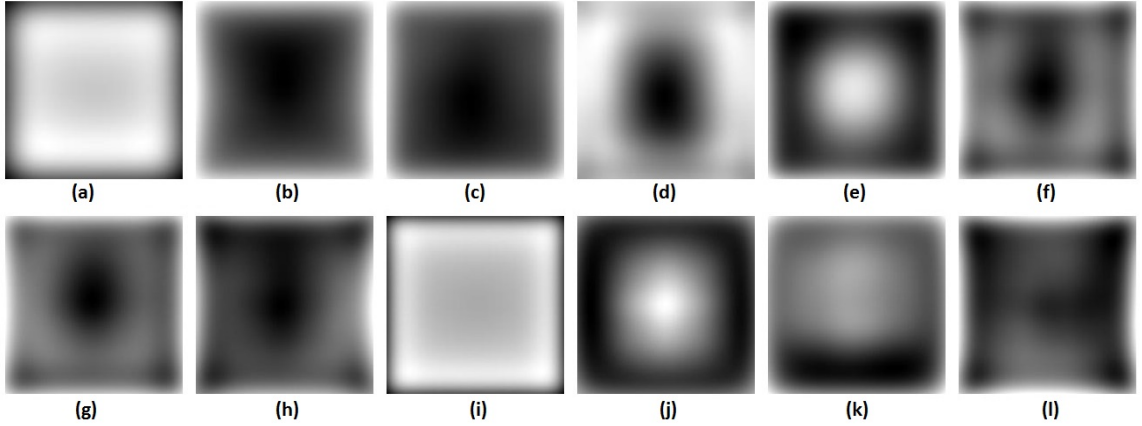


Figure 3.2: Spatial bias map of algorithms on Koehler *et al.* [29] dataset (a) Torralba [49], (b) HouCVPR [16], (c) HouNIPS [17], (d) AWS [13], (e) Itti-CIO2 [19], (f) ImageSignatureLab [15], (g) ImageSignatureRGB [15], (h) SDSR [46], (i) AIM [8], (j) GBVS [14], (k) Yan [56] and (l) SOC [40].

figure, high values within the center region (e.g. GVBS) imply strong relative central bias inherent in algorithm output and conversely, some algorithms (e.g. AWS) exhibit a relative peripheral bias. This provides a useful foundation for analysing differences in model performance across different types of data and also has implications for both standard and shuffled ROC scores. In particular, a strong relative center bias carries a predictable benefit for the standard ROC metric, however a stronger relative peripheral bias carries a benefit for the shuffled AUC metric.

3.2.2 Removal of center bias

If a saliency algorithm carries a strong center bias, then significantly higher values within the center region will appear consistently across all output maps. As stated before, because of the presence of center bias, the evaluation of saliency algorithms may be misleading. The term center bias removal technique refers to the strategies

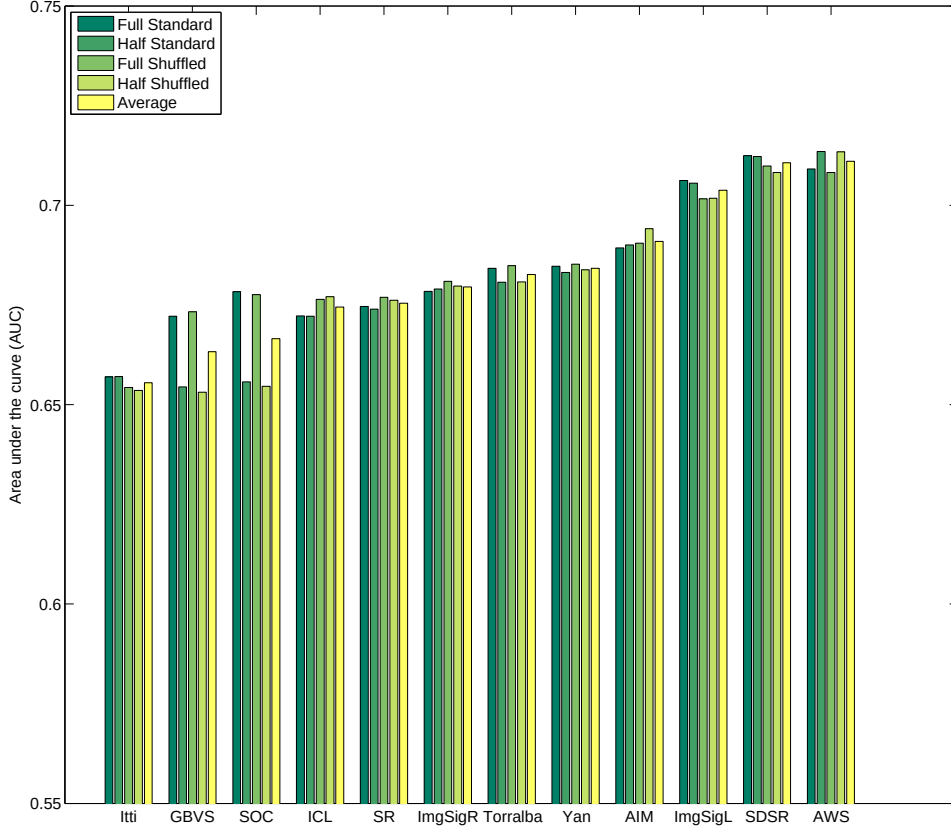


Figure 3.3: Rank order based center bias removal on Toronto dataset

we have applied to remove the effect of center bias within different saliency maps. These techniques make the statistics of saliency values uniform at every pixel location across all saliency maps. In this thesis, we consider two options to remove center bias in algorithm output to diminish sensitivity to spatial bias. In the current work, we propose two different processes of center bias removal that derives from the computational structure of the saliency models themselves.

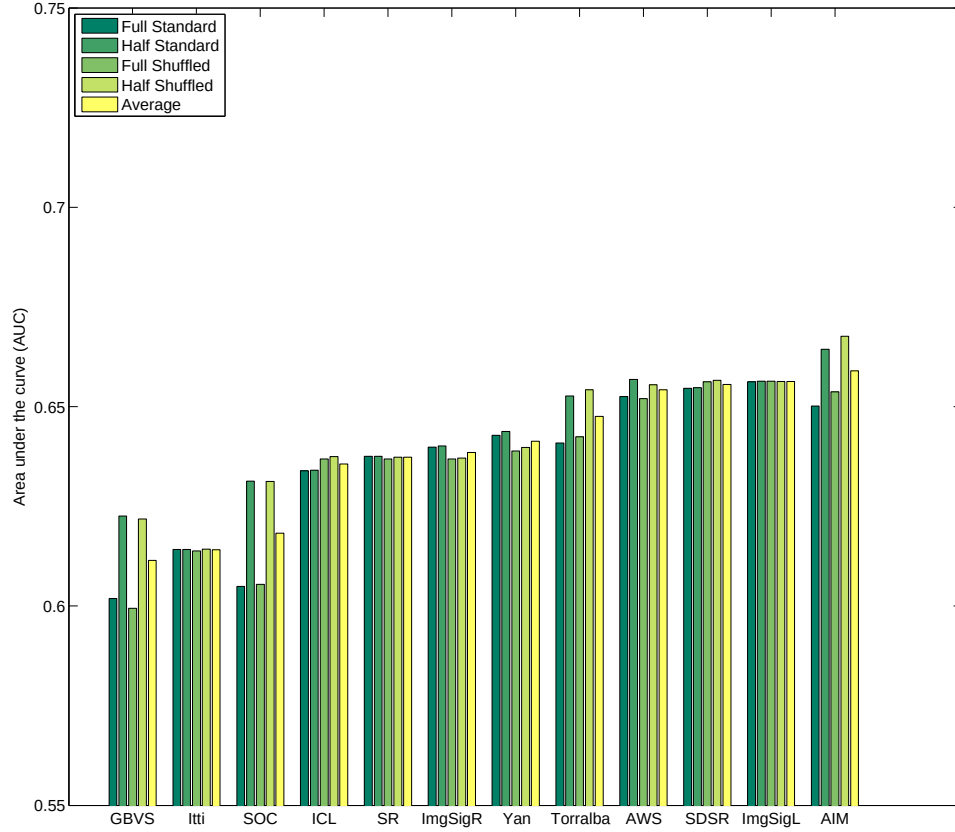


Figure 3.4: Rank order based center bias removal on Judd dataset

Rank order based center bias removal

To produce a uniform spatial distribution for the output of saliency algorithms, we wish to normalize the saliency values across all output maps on a per-pixel basis. We achieve this through the rank ordering of values in ascending order for a given pixel location across all saliency maps for the entire dataset. Suppose $M_i(x, y)$ is the pixel value corresponding to coordinate (x, y) in i th saliency map produced by any

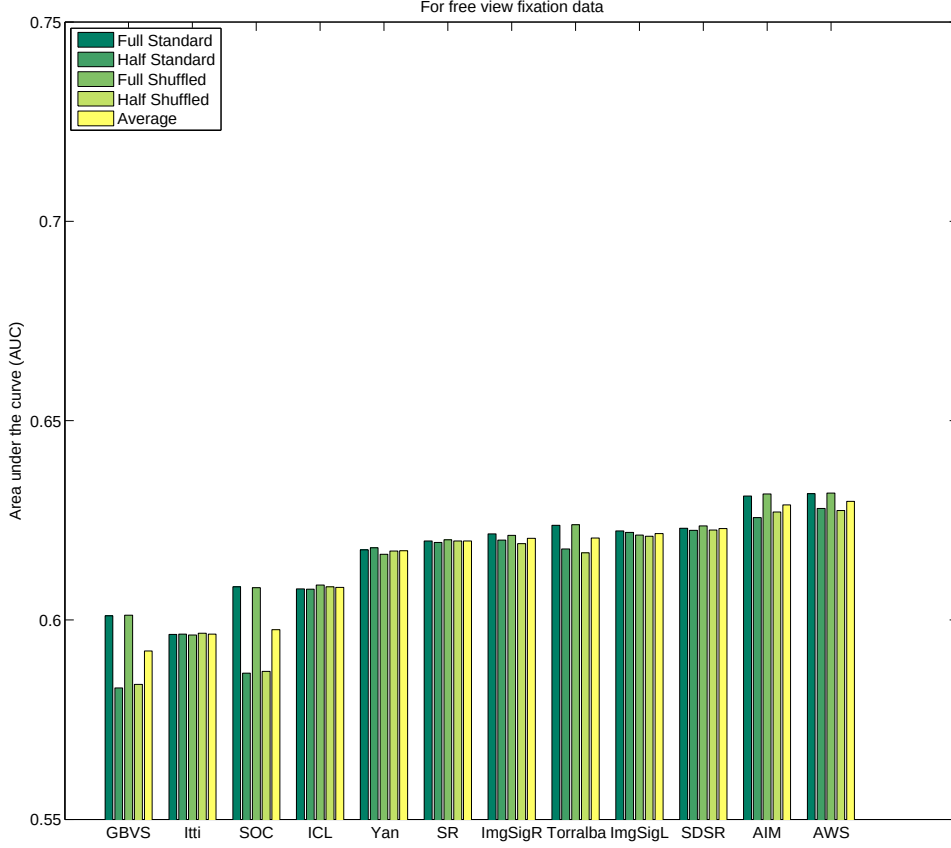


Figure 3.5: Rank order based center bias removal on Koehler *et al.* [29] dataset

algorithm where $i = 1, 2, 3, \dots, N$, and N is the total number of images considered for each algorithm. Now, for each location (x, y) we calculate the rank order of the N saliency values in M_i . Subsequently, values of $\frac{1}{N}, \frac{2}{N}, \frac{3}{N}, \dots, 1$ are assigned sequentially to the saliency maps based on this rank order. This imposes a uniform distribution of scores with the range $\frac{1}{N}$ to 1 across all maps, and for each pixel location. Following this center bias removal from all images of the dataset, the sum of all values corresponding to a given pixel location across all saliency maps is $\sum_{i=1}^N M_i(x, y) = \frac{N+1}{2}$. After spatial

center bias removal, saliency output may be evaluated using both standard ROC [23] and shuffled ROC [4; 57] analysis. The rankings of algorithms after removal of spatial bias in output on Toronto [8], Judd *et al.* [23] and Koehler *et al.* [29] datasets are shown in Figures 3.3, 3.4 and 3.5.

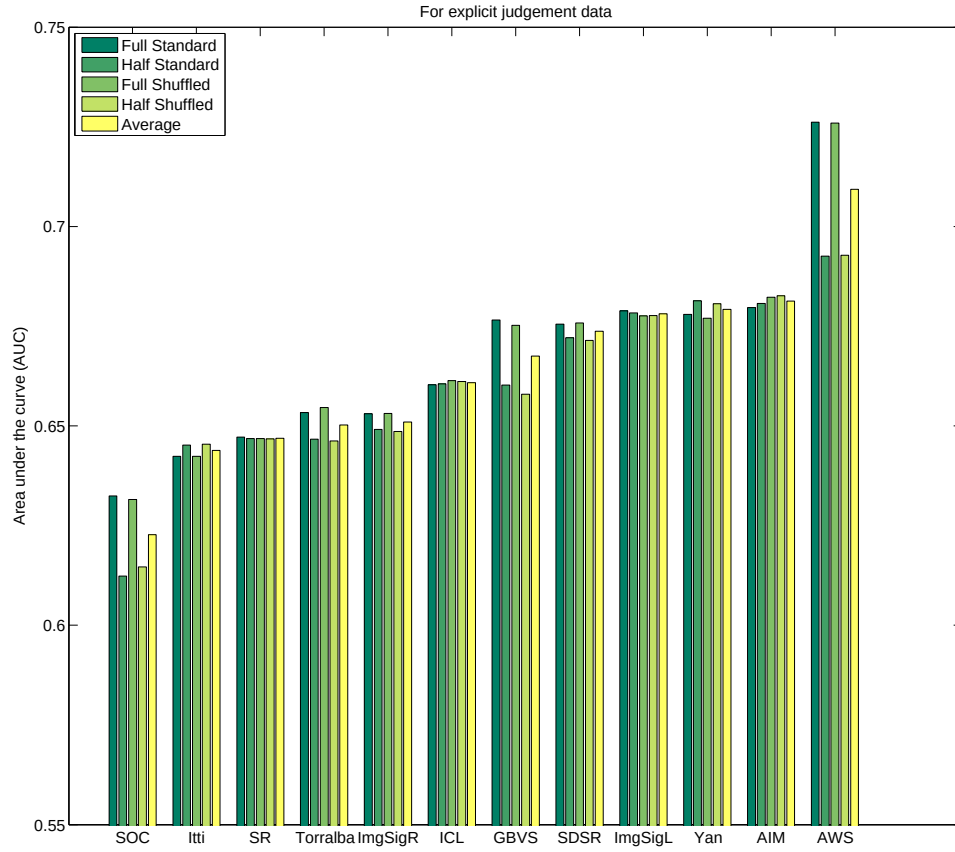


Figure 3.6: Rank order based center bias removal on Koehler *et al.* [29] dataset (Explicit Judgment)

Whitening based center bias removal

An alternative to rank ordering, is to apply a more standard statistical whitening operation. First, individual saliency maps are whitened to produce maps with zero mean and unitary standard deviation. Subsequently, pixel-wise whitening is carried out across the saliency maps for each individual pixel location. After this operation $\sum_{i=1}^N M_i(x, y) = 0$ and each algorithm carries a 0 mean and unit variance across all saliency maps for each pixel location. The ranking of algorithms after center bias removal using this process on Toronto [8], Judd *et al.* [23] and Koehler *et al.* [29] datasets are shown in Figures 3.7, 3.8 and 3.9.

A few observations related to both strategies for center bias removal are as follows:

- Resizing the input images changes the scale space spanned by the features for some algorithms, which may alter performance. Sensitivity to scale is a factor that is important to control for outside of spatial bias.
- Both standard AUC analysis [23] and shuffled AUC analysis [57] produce very similar results for a fixed image scale with normalized outputs.
- In contrast to the standard ROC benchmarks, spatial bias removal for algorithm output produces more consistent performance of algorithms across different datasets, and metrics.

3.2.3 Discussion

To address central tendency of human fixation data, different algorithms apply different mechanisms to consider this phenomena into the saliency detection process.

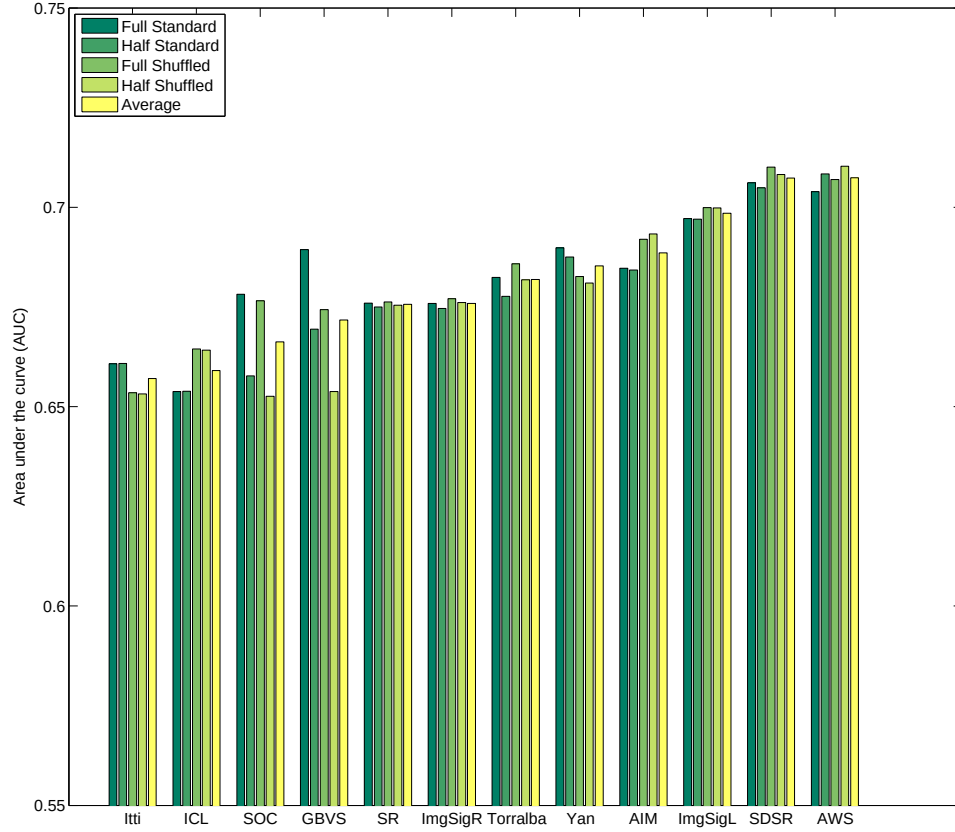


Figure 3.7: Whitening based center bias removal on Toronto dataset

For example, those mechanisms include post processing Gaussian blurring with different kernel size and/or normalization of the pixel with respect to surrounding neighbor pixels. As the extent of center bias is different from algorithm to algorithm, thus we need to know the degree to which algorithms consider the bias and the process to remove it. In this section, we first discussed the extent of center bias presented in saliency maps. The visualization in Figure 3.2 shows the relative bias of algorithms towards specific spatial regions. We also discuss two possible ways of removing center

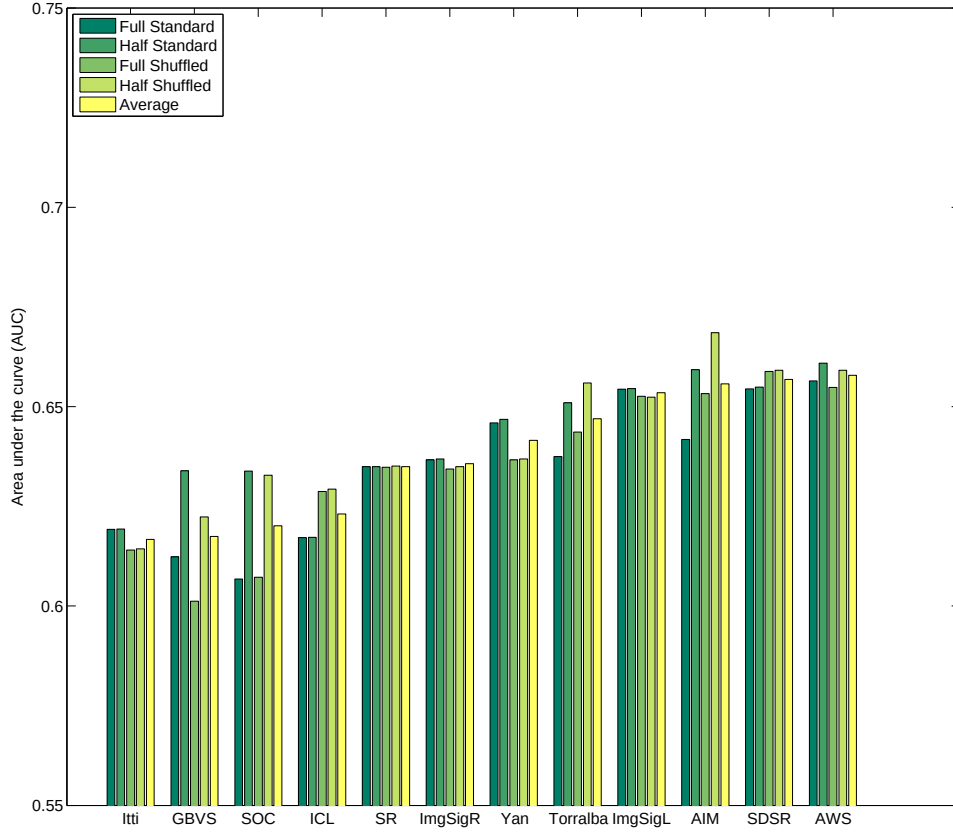


Figure 3.8: Whitening based center bias removal on Judd dataset

bias from algorithms. Furthermore, we perform algorithm benchmarking experiments on center biased removed saliency maps of different algorithms. We find relatively consistent ranking of saliency algorithms across different data sets and tasks.

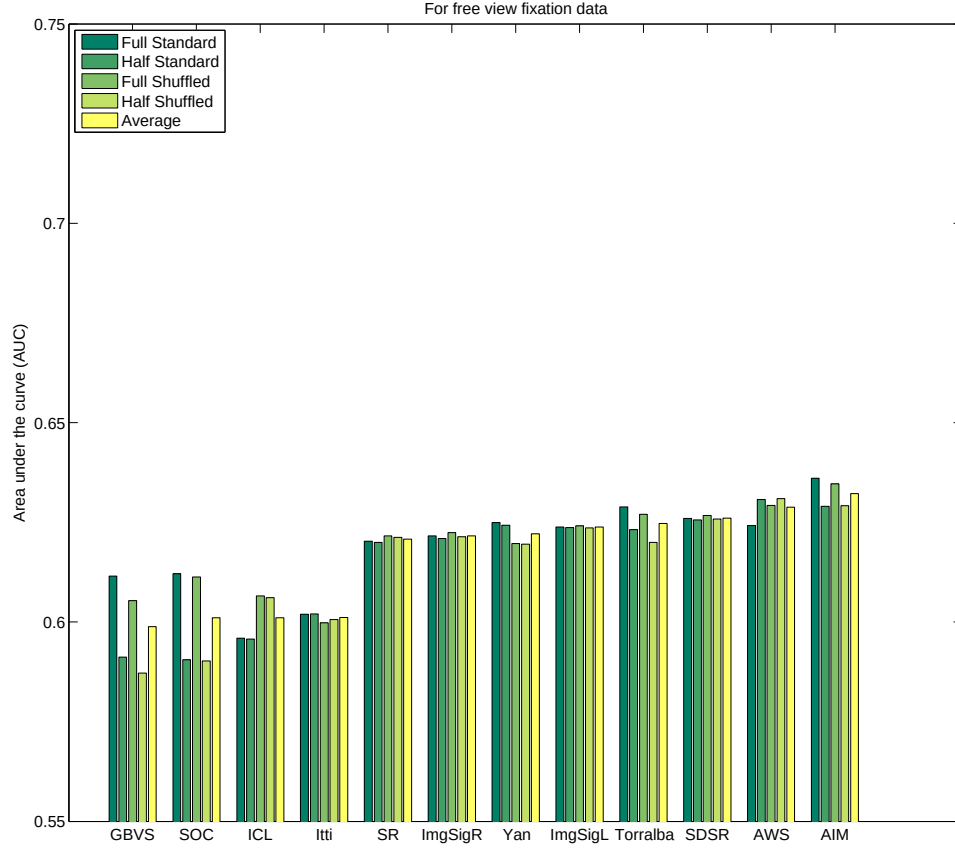


Figure 3.9: Whitening based center bias removal on Koehler *et al.* [29] dataset

3.3 Explicit judgment

3.3.1 Benchmarking of explicit judgment

In viewing an image, an observer will fixate on locations in the image by their own volition, due to the nature of stimulus patterns or other external factors. Rather than considering fixated locations, an alternative is having participants in an user study select a specific location of interest. This presents an additional quantity of value

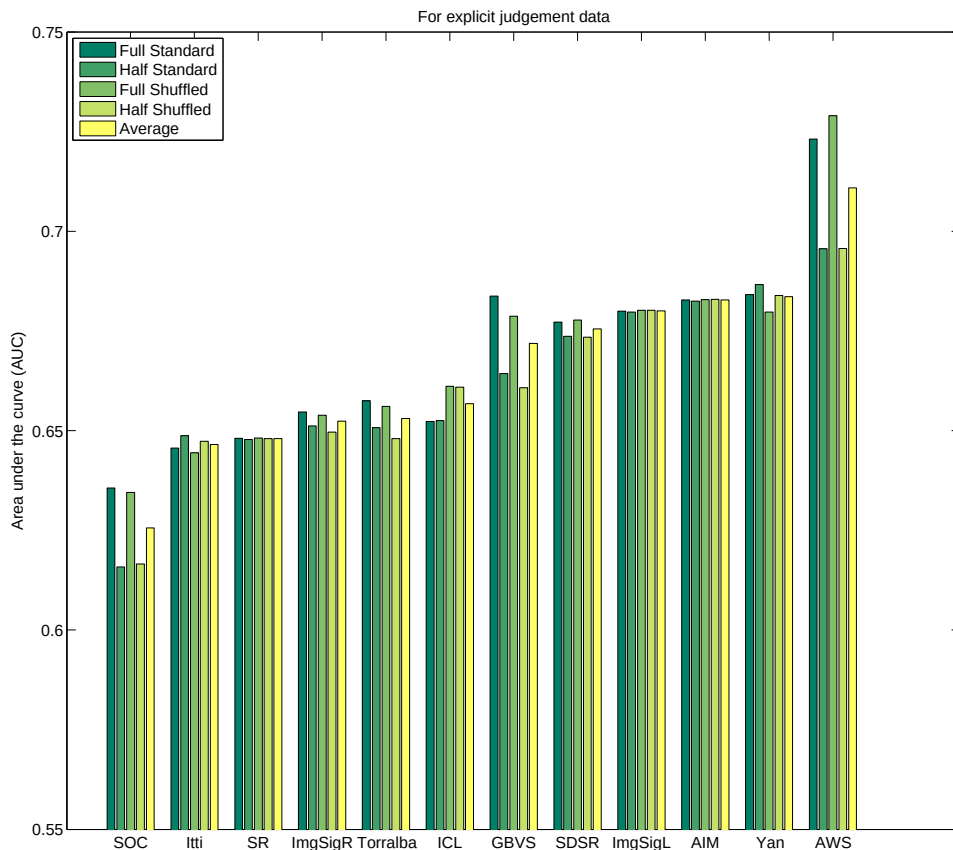


Figure 3.10: Whitening based center bias removal on Koehler *et al.* [29] dataset (Explicit Judgment)

for saliency algorithms to predict. This final selected location is called the explicit judgment of the observer corresponding to the most salient location. Saliency models typically attempt to predict the gaze points in fixation data rather than explicit judgments. The recent study of Koehler *et al.* [29] that addresses this shortcoming, provides a rich dataset containing explicit judgments for 100 human observers. Apart from the explicit judgment task, Koehler *et al.* also considered three other tasks (free viewing, saliency viewing and a cued object search task). The free viewing

Table 3.3: Benchmarking performance for the Koehler *et al.* [29] dataset

Task	Free Viewing		Object Search		Saliency Viewing		Explicit Judgment	
AUC	Stand.	Shuf.	Stand.	Shuf.	Stand.	Shuff.	Stand.	Shuf.
HouCVPR [16]	.726	.639	.709	.632	.735	.652	.735	.657
SOC [40]	.786	.632	.722	.630	.742	.643	.736	.643
HouNIPS [17]	.719	.626	.700	.614	.733	.644	.748	.674
ImgSignRGB [15]	.739	.639	.725	.634	.750	.652	.749	.657
Itti-CIO2 [19]	.753	.629	.743	.638	.754	.638	.758	.654
Torralba [49]	.769	.635	.749	.629	.766	.642	.763	.652
SDSR [46]	.747	.643	.727	.635	.756	.657	.768	.683
ImgSignLab [15]	.746	.642	.727	.633	.756	.659	.773	.688
AIM [8]	.753	.632	.733	.624	.756	.640	.774	.665
AWS [13]	.742	.645	.719	.632	.709	.664	.778	.699
GBVS [14]	.782	.633	.772	.649	.776	.635	.781	.655
Yan [56]	.769	.644	.753	.647	.775	.656	.782	.682
IOC score	.847	.716	.875	.770	.847	.724	-	-

task represents searching behavior without any particular goal, which is the scenario traditionally considered in most datasets in the literature. In the saliency search task, observers were asked to decide whether the right or left portion of the image is more salient than the other. Finally, in the object search task, observers were instructed to find some pre-defined objects. Except for the explicit judgment case, each of these tasks is examined through eye tracking experiments.

Relative to fixation data, very little evaluation has focused on the ability of saliency algorithms to approximate locations judged as salient in an explicit judgment task. Although Koehler *et al.* [29] presented results in this regard, only a very narrow range of algorithms were considered, and therefore there is value in considering a wider range of established algorithms. In Table 3.3 we have reported benchmark results for 12 algorithms including AUC scores from ROC analysis for all four tasks: free viewing, saliency viewing, cued object search and explicit judgment tasks. In AUC

based evaluation, we have considered two different types of ROC analysis: Standard ROC [47; 23] and shuffled ROC [4; 57] that each apply a signal detection approach to evaluate the prediction performance of saliency maps, but that treat spatial bias differently. As may be seen from Table 3.3, there is some consistency in the performance of algorithms, but also significant variability dependent on the specific nature of ROC analysis. In results that follow, we have shown that both standard and shuffled ROC analysis may be affected by different types of data bias. To compare the results with the upper bound of prediction [31], we have also reported the inter-observer congruency (IOC) score of eyetracking data. As explicit judgment locations are not collected by gaze locations, there is no IOC score for this case.

3.3.2 Approximation of explicit judgments

Gaze patterns in the absence of an explicit task (free viewing) and with a prior task (object search, saliency viewing) are both experimental paradigms that have been widely studied in the literature [26; 23]. In contrast, the more direct process of making a manual selection proposed by Koehler *et al.* [29] diverges from the traditional methods for examining the visual selection process through eye tracking. Koehler *et al.* argued that algorithmic determination of saliency bears a closer resemblance to locations selected through explicit judgment than fixation data collected from any of the other three tasks. In the current work, this suggestion is confirmed in Table 3.3, with further information on the performance landscape for different popular algorithms across the different tasks.

We believe that improving the prediction of explicit judgments is likely more pru-

dent than improvements to performance for the traditional fixation tasks for several reasons: Explicit judgment captures the most salient locations within a scene, through a selection process that is less clouded by noise from spatial bias and fixation mechanisms, and with more relation to content relevant to the role of saliency as a front-end selection mechanism in computer vision. Although there is evidently value in focusing on explicit judgment prediction, this has not been on the radar of development in algorithms targeting visual saliency, and also there is a relative paucity of this type of data. Moreover, in the future the tradition of examining perception through eye tracking across many areas of study is likely to continue. In this subsection, we therefore strive to present a means of simulating explicit judgment data via saliency and eye tracking data. This allows for broader capability in the qualitative analysis that explicit judgment provides, and also as a means of providing a larger corpus of simulated explicit judgment data to drive improvements in visual saliency prediction.

3.3.3 Comparison of tasks in terms of IOC

The viewing pattern of observers varies from task to task. Being motivated by bottom up saliency, the free viewing task is meant to capture the task-independent bias in viewing the image. All fixation data also carries an implicit tendency to look at the center of the scene [47]. Such center bias is present even when stimuli are not placed at the center of the image. Unlike the free viewing task, when contextual guidance is imposed, then fixated locations more closely follow the regions where target objects are present rather than and center bias is somewhat diminished [57]. This scenario occurs to a significant extent for object searching task and also occurs

to a lesser extent for the saliency viewing task. Because of the differences in viewing pattern, the agreement among the observers varies from one task to another. To observe this variation, we have calculated the correlation among the inter observer congruency (IOC) scores of different types of task presented in Koehler *et al.* [29] dataset. Traditionally, one IOC score is calculated for an entire dataset [31] which is presented as human observer performance in Table 3.3. To examine this quantity in greater detail, we have calculated IOC scores for individual images and found correlation for IOC scores among different tasks on a per-image basis. The correlation between free view vs. object search, free view vs. saliency view and object search vs. saliency view are 0.28, 0.58 and 0.29 respectively. The scatter plot of these measurements is presented in Figure 3.11. These results suggest that the agreement in viewing pattern across observers for free viewing correlates with agreement for saliency viewing to a greater extent than tasks. This also suggests overlap in the underlying factors driving gaze selection for these two conditions. In the object search task, the distinction between bottom-up and top-down factors in viewing evidently plays a greater role, lending some credence to the bottom-up or stimulus driven claim that is typically attached to free viewing fixation data. With that said, it is evident that there are significant differences between any of the fixation based measurements and the explicit judgment condition, and we shed further light on this point in the sections that follow.

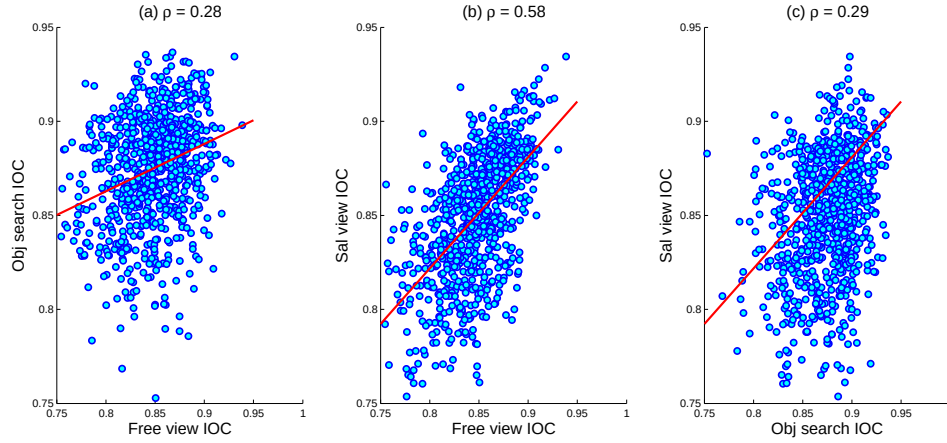


Figure 3.11: Scatter plot showing the correlation among the IOC of free view, obj. search and sal. view tasks

3.3.4 Relation of explicit judgment to IOC in fixation tasks

If there is strong agreement in viewing pattern across observers for fixation data across different tasks, one might intuit that overlapping locations may be especially salient. One might also suspect that such locations may be among those selected in the explicit judgment task. Therefore, we hypothesize that a high IOC score for an image may be indicative of instances wherein a fixation map makes a good approximation of explicit judgment. To test this hypothesis, we have calculated correlation between the auROC performance of the fixation maps in predicting locations selected via explicit judgment, and the corresponding IOC score for a given image within-task. The scatter plot is shown in Figure 3.12. We find that the correlation performance for the object search task is lower than that of free view and saliency view tasks. In conjunction with the results reported in Table 3.4, this analysis hints at the relatedness of the different task conditions. Evidently, there is some top-down influence within each of these tasks, and influence from cognitive processes unrelated to saliency. However,

that algorithms bear a stronger resemblance to explicit judgment data when the IOC score is high strongly suggests that these processes converge, and more of the computation is stimulus driven, or bottom-up.

We know that viewing pattern of observers is dependent on both top-down and bottom up influences. Top down influence is greater in object searching task whereas bottom up influence dominates in free viewing and saliency viewing task. Because of top-down influence people used to perform a guided search based on observers experience or pre-instruction regarding task. In most cases, people normally look at complex scenes having multiple objects or salient location. When observers view a complex scene, based on their previous knowledge they fixate in different locations. So the combined effect of definite target and observers past experience drive observers to variable attention behavior. For the same reason IOC scores for those images and the approximation performance of algorithms become unpredictable. On the other hand, no external influence affects the viewing behaviour for the case of bottom up saliency. Irrespective of users experience, free view or saliency view task follow similar viewing response which is enough to correlate the algorithm performance and IOC of image. However, top down effect does not sustain for long time because when a observer achieves the pre-instructed goal of searching task, he switches his search task to free view mode. In general, when searching task is more likely to free viewing saliency driven, the viewing pattern become similar across observers intuitively.

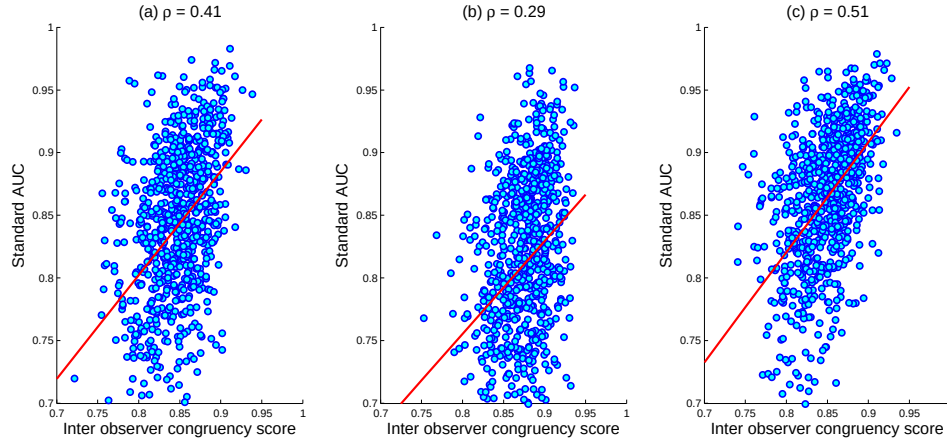


Figure 3.12: Scatter plot showing the correlation between Standard AUC of explicit judgment approximation vs. IOC scores of (a) free view (b) object search (c) saliency view fixation data

3.3.5 Explicit judgment prediction model

In this subsection, we focus on prediction of explicit judgment. There are two goals that are central to this section. The first goal is to determine how well existing models of visual saliency are able to predict locations selected through explicit judgment. While some sense of this is already provided in our benchmarking results, we further examine the strength of predictions that may be achieved through an ensemble approach that relies on existing saliency algorithms. This provides a new standard for future efforts in prediction of explicit judgment data, but also informs on how significant the difference is between models tuned to perform well on fixation data, and performance that is possible for explicit judgments. A second important goal of this section is to relate explicit judgment data to fixation data through a predictive model. Fixation data is ubiquitous, while measurement of explicit judgments of visual saliency is relatively rare. These two types of experimental data offer different

vantage points on the saliency of image content to a human observer. There is value to each type of data in differences in qualitative analysis that is possible through heatmaps or other visualizations, and quantitative analysis in benchmarking and predictive models. We therefore examine the extent to which existing eye tracking data can be used to simulate the distribution that one would expect from an explicit judgment experiment. Given that both types of data are available within the Koehler *et al.* [29] dataset, this also allows for the relationship between the two types of data to be examined.

A goal of prediction in this context, is to establish the extent to which explicit judgment prediction may be improved beyond existing standard saliency models while relying on similar principles and features. From the benchmarking results in Table 3.3, we know that individual saliency algorithms can approximate explicit judgment to a certain degree. If we constrain our own prediction model to the space of features, and saliency measures spanned by existing models of visual saliency, this provides an indication of how well principles driving existing saliency models translate to prediction of explicit judgment data. This evaluation also succeeds in setting an improved benchmark score for explicit judgment prediction. To achieve these goals, an ensemble prediction based on a range of saliency models is used to make this prediction.

Table 3.4: Prediction of explicit judgment using Gaussian blurred free view, object search and saliency viewing fixation map respectively.

Blur level	1	2	3	4	5	6	7
Sal. View	.78	.86	.87	.86	.85	.85	.84
Free View	.76	.84	.84	.84	.84	.82	.81
Obj. Search	.74	.81	.81	.81	.79	.79	.79

Before delving directly into performance of our ensemble classifier for explicit judgment prediction, it is useful to further consider task relatedness for fixation tasks, and explicit judgments. If data from any of the fixation conditions is converted to a continuous density map via Gaussian blurring (as is standard in saliency evaluation), performance in predicting explicit judgments varies as a function of the degree of blur. Approximation of explicit judgment locations by fixation data from the three different fixation tasks is shown in Table 3.4. Both free viewing and saliency viewing fixation maps can approximate explicit judgment quite well, and better than object search fixations. In the object search task, although observers are directed to find a pre-defined object, the wide variety of objects presented in a typical scene influences fixations on every salient items in the scene which may not be task relevant. The relationship to explicit judgment indicates that there is indeed a significant degree of task-independent commonality between free viewing, saliency viewing and explicit judgment.

We have also examined how the results appearing in Table 3.4 interact with IOC on a per-image basis. To accomplish this we compare the base IOC scores (per image) within each category of fixation data with the auROC performance in predicting explicit judgments from the fixation data. This provides an understanding of the relationship between IOC within each type of fixation data, and the relatedness of fixation and explicit judgment data as a function of IOC. This is measured by examining the Pearson correlation on a per image basis between image IOC for the fixation data, and the auROC for the prediction of explicit judgments from fixations. Correlation values from this line of experimentation are as follows:

- Free viewing IOC vs. auROC for prediction of explicit judgment from free viewing data: 0.41
- Object search IOC vs. auROC for prediction of explicit judgment from object search data: 0.29
- Saliency viewing IOC vs. auROC for prediction of explicit judgment from saliency viewing data: 0.51

The ability to predict explicit judgment based on fixations is good when the IOC for the fixation data associated with the same image is high. Statistically, this result indicates that when observers exhibit more similarity in their fixations for a particular image, then factors that drive the selection of fixated locations and explicit judgments also become less disparate. This hints at overlap in factors driving the two processes, and also that per-image confidence for simulating explicit judgment data from fixations, might be measured through the associated IOC score.

As mentioned, a second goal in this section is the simulation of explicit judgment data when only eye tracking data is available. From Table 3.4, it is evident that fixation data is also indicative of locations selected in the explicit judgment task. Thus, it is natural to consider the combined strength of fixation data and saliency algorithms in simulating a measure of explicit judgment. In line with this goal, we consider only free viewing fixation data given that this tends to be the standard data type that is widely available.

To train the prediction model, the image set with explicit judgment ground truth is divided into training and test sets (half/half). For each image, there are 100 locations

selected as most salient through explicit judgment and these are labelled as positive samples. From each image, we also randomly sample locations not among the positive samples as negative samples. Output values from 12 saliency maps for the positive and negative pixel locations are sampled as training data. In addition, in support of the simulation of explicit judgment, fixation density from blurred fixation maps are sampled as an additional statistic. Given that there may also be dependency on spatial location, and also on the spatial dispersion of positive samples across task, we also include the (x,y) coordinates of samples as additional statistics. Features from all the images, are used to train an ensemble of bagged decision trees for regression (Random Forest [6]). For testing, saliency values and fixation densities for each pixel location are then used to predict explicit judgments. Standard and shuffled AUC scores, subject to a varying number of trees are shown in Figure 3.13 (a) and (b). Figure 3.13 includes the decomposition into different combinations of feature types, including only saliency output, only fixation data, and combinations thereof. Note that center bias removal is not applied here as most of the cases explicitly include spatial position as a feature. This illustrates the extent to which existing saliency algorithms are capable of predicting locations selected in explicit judgment, and also addresses the extent to which explicit judgments may be simulated from fixation data with saliency algorithms as an adjunct source of information. Figure 3.14, demonstrates model output that is possible under different conditions or through different combinations of feature types.

As a baseline for assessing the boosted classification performance of saliency algorithms in predicting explicit judgment locations, it is useful to also examine how

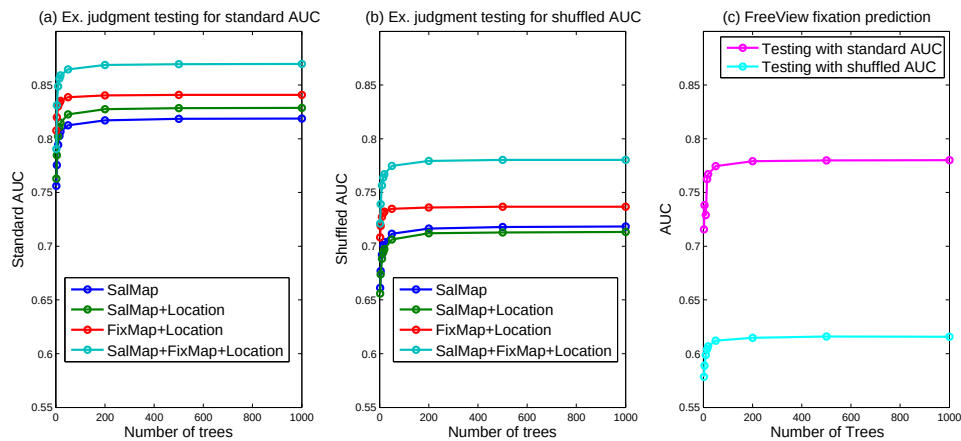


Figure 3.13: Performance for explicit judgment prediction. (Best viewed in digital format)

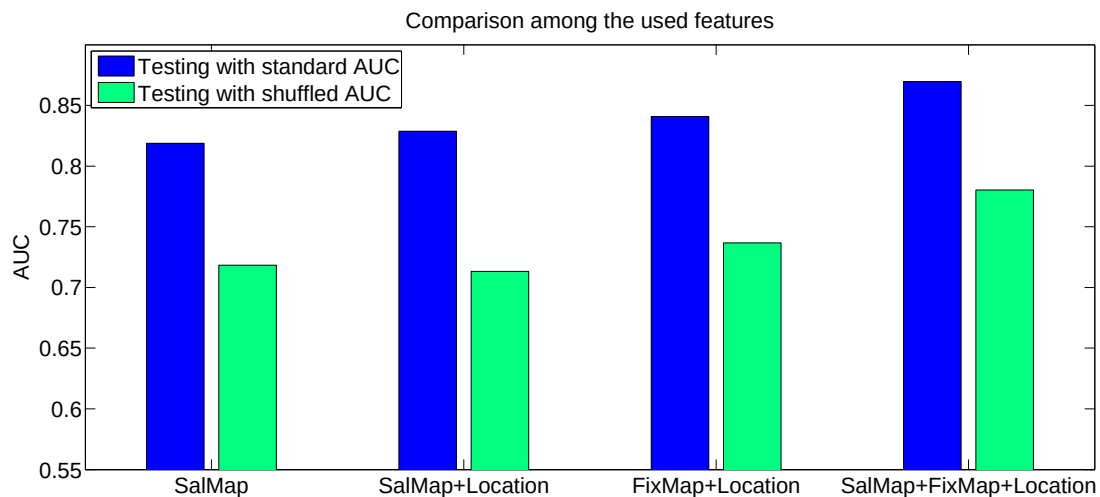


Figure 3.14: Explicit judgment prediction using different feature sets.

much a boosting approach improves performance for the traditional fixation based evaluation. Experiments measuring this case follow the same process described for explicit judgment prediction, in using ensemble performance across saliency algorithms to predict fixated locations. Results relating to these experiments appear in Figure 3.13 (c). We find that although a boosted classifier based on saliency algorithms is

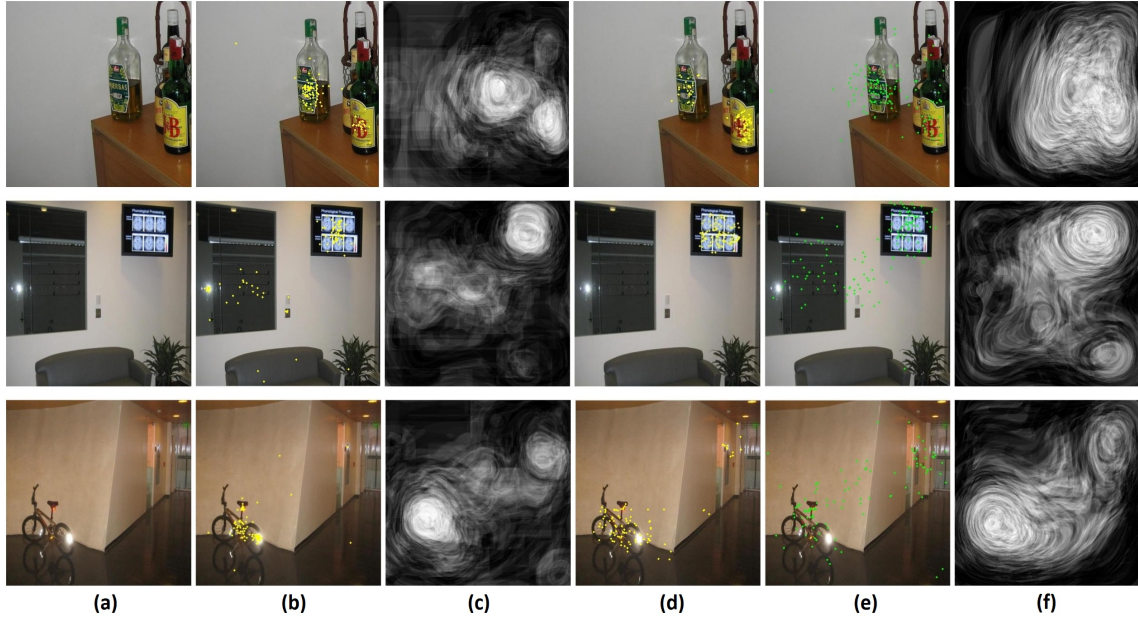


Figure 3.15: Sample output column (a) input image (b) ex. judgment locations (c) predicted ex. judgment map (d) simulated ex. judgment locations (e) free view fixation locations (f) predicted fixation map

able to significantly improve the performance of explicit judgment prediction, there are no significant gains in fixation data prediction (refer to Table 3.3 to make this comparison). This suggests that similar factors are involved in determining explicit judgment locations and fixations, however models have been more finely tuned to characteristics of fixation data.

3.3.6 Explicit judgment performance prediction

A sample of input images and their associated actual explicit judgments, predicted explicit judgments (saliency output) and simulated explicit judgment data are presented in Figure 3.15. This also includes the true fixation data and predicted fixations (saliency) for comparison.

Fixation data from a free viewing task is the most common form of data in evaluating visual saliency. However, there are evidently benefits to having the ability to leverage explicit judgment data in guiding algorithm development and assessing performance. In this section, we assess whether the prediction of explicit judgment made possible by our model is of sufficient quality that this output may be used in a surrogate role for quantitative assessment. For example, given a new algorithm for visual saliency prediction, and various fixation data sets, can one predict the relative performance of the new algorithm for the explicit judgment task when only fixation data is available.

The predicted explicit judgment map consists of a topological representation of expected explicit judgment locations. Evaluation, whether by fixations or explicit judgment typically relies on binary values corresponding to discrete pixel locations. To generate this type of representation for quantitative evaluation on simulated explicit judgment data, we first normalize the predicted explicit judgment map (saliency output). Subsequently, 100 randomly sample locations are chosen based on the values in the predicted explicit judgment map, ignoring locations where the value within the explicit judgment map falls below a set threshold. This strategy was found to produce superior results to non-thresholded sampling from the predicted explicit judgment map. Given discrete coordinates for simulated explicit judgment data, evaluation may proceed based on the standard methods for ROC analysis.

Saliency algorithms were evaluated based on free viewing fixation data, explicit judgment data, and simulated explicit judgment data. The critical factor in this analysis is the relation of performance on true explicit judgment data to simulated

explicit judgment data. The correlation between ROC scores of different algorithms for true explicit judgment data, and simulated explicit judgment data is 0.92. In contrast, correlation between per image ROC scores in predicting fixation data and in predicting explicit judgments is 0.46. This provides confidence that the simulated explicit judgment data derived from fixation data may present a suitable approximation for analyzing relative algorithm performance for explicit judgment tasks when only eye tracking data is available, and that saliency is necessary to provide a bridge between these disparate sources of data.

3.3.7 Details of post-processing blur for ensemble prediction

As with most benchmarking efforts in visual saliency prediction, performance is affected by the amount post-processing blur. This effect is presented in Table 3.5.

Table 3.5: For a fixed NTree = 15, the effect of different Gaussian blur level in the explicit judgment prediction model

Blur level	1	2	3	4	5	6	7
Standard AUC	.83	.85	.85	.86	.85	.85	0.84
Shuffled AUC	.73	.74	.76	.76	.76	.75	.73

3.3.8 Saliency and Segmentation

Given that explicit judgment locations are likely to be under greater cognitive control than early fixations, explicit judgment location may arguably provide a stronger marker for object locations than fixation data. To evaluate this hypothesis, we have carried out experiments involving object segmentation performance based on both predicted explicit judgment locations and predicted fixations using 200 images in-

cluding one and two object cases from the SED dataset [2]. To evaluate segmentation performance, we have calculated the F-measure by the following equation:

$$\text{F-measure} = \frac{\text{precision} \times \text{recall}}{0.5 \times (\text{precision} + \text{recall})}$$

F-measure scores for the overall dataset are shown in Table 3.6 and the corresponding precision-recall curves in Figure 3.16. These results indicate that predicted explicit judgment maps are more successful than predicted fixation maps in highlighting segmented object locations within a scene. Moreover, the performance improvement using predicted explicit judgment maps over predicted fixation maps in the two object case is higher than that of one object. This also suggests that the distinction between predicted explicit judgment maps vs. fixation maps becomes more prominent for complex scenes having multiple objects.

Table 3.6: F-measure of segmentation performance

Object case	Exp. judgment map	Fixation map
One object	0.698	0.689
Two object	0.604	0.566

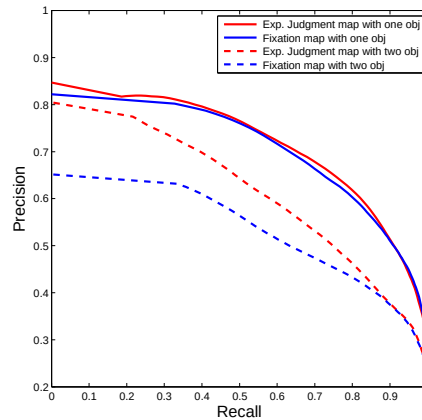


Figure 3.16: Precision Recall Curve for one and two object case using both explicit judgment and fixation map

3.3.9 Discussion

In this section, we address the problem of computational modeling of visual saliency in considering explicit judgment data, as opposed to the more common alternative of predicting fixations within human gaze data. There are a number of contributions and results of importance that are derived from this investigation:

1. Benchmarking results are presented for several algorithms across four different tasks. This demonstrates the performance landscape of popular algorithms for visual saliency for explicit judgment of salient locations. We also motivate why this alternative form of ground truth may be advantageous for perception related predictions and computer vision applications.
2. The experimentation included also provides an indication of task relatedness. Free viewing and especially saliency viewing appear to be driven by factors that overlap with explicit judgment. With that said, there are different confounding factors associated with each type of data, and reason to believe that explicit judgment is more proximal to representing the quality that saliency algorithms aim to predict.
3. Results indicate that IOC predicts the similarity of fixation data to explicit judgment data. This implies that IOC may be a reliable measure for gauging the extent to which vision is stimulus driven, and how well explicit judgment may be inferred from fixations.
4. Existing algorithms designed for fixation prediction do reasonably well in predicting explicit judgments. However, these algorithms perform much better

when tuned to predict this data specifically. This implies that underlying principles in existing saliency algorithms are suitable for predicting explicit judgment, but that some optimizations of algorithm performance have been specific to statistical properties of the spatial layout of fixation data.

5. Explicit judgment marks an alternative window into perceptually important content to human observers. While eye tracking data is relatively standard in many research domains, explicit judgment tasks are not. We have also presented a benchmarking result that surpasses current saliency algorithms for explicit judgment prediction, in addition to a method that allows for fixation data and saliency models to approximate explicit judgment data to an extent that qualitative interpretation and quantitative analysis towards explicit judgment prediction may be carried out when only fixation data is available.

As a whole, this work establishes performance standards for the explicit judgment task, introduces an alternative ROC based benchmark strategy, demonstrates the relatedness of perceptual tasks, and provides methods for data simulation across different perceptual tasks.

Chapter 4

Digital Media Rating

4.1 Inter-observer congruency prediction

If an algorithm for predicting visual saliency performs poorly for a particular image, this may instead be due to the properties of the image itself rather than a failure of the algorithm [31]. As mentioned, there exists an upper bound on performance dictated by agreement in viewing patterns among observers, and quantified by the Inter-observer Congruency score. Therefore, the performance of a saliency detection algorithm critically depends on IOC of the input image. There are several factors which may be responsible for a low IOC score. Firstly, IOC can be low if the image has no *salient* content i.e. nothing prominent or especially significant to attract attention. This is a scenario that is not uncommon for naturalistic images for which algorithm performance and upper bounds dictated by IOC are especially low [40]. Secondly, IOC can be low if the image has many salient locations, coupled with a shorter viewing time. Such scenarios may arise when the image contains many

objects or outliers with potential to attract attention in a bottom-up fashion. It is also commonly observed that the first few fixation points show significant congruency among observers with inter-observer agreement progressively dropping off with time [31]. On a slightly longer time course, task dependence and contextual guidance exert greater influence on search behaviour which also interacts with congruency. Given a wide array of salient targets in a single scene, fixations are divided differently among observers across targets. In Figure 4.1, we show a ranking of images corresponding to the 5 highest and 5 lowest IOC values for the three datasets considered in the current work. There are a few observations that may be made immediately at an observational level concerning these images. Those that exhibit high IOC scores tend to be marked by one of: (i) A small number of discrete, or localized regions of high contrast (ii) the presence of people, or faces in the scene. (iii) the presence of text within the scene. On the low end, images tend to consist of a broad distribution of structure, cluttered or busy scenes, and landscapes. The low IOC images also appear somewhat more generic in what they capture.

4.1.1 IOC computation

IOC is typically quantified based the average IOC across all images within a dataset, but in the current study we consider the IOC score corresponding to single images. For 3 individual and distinctive datasets, we have computed individual IOC scores for each image. This is accomplished through the *leave one out* method for IOC computation [51]. Each dataset involves eye tracking data from n observers. For any $i \in n$, a predictive map based on $(n - 1)$ observers is used to predict fixation data

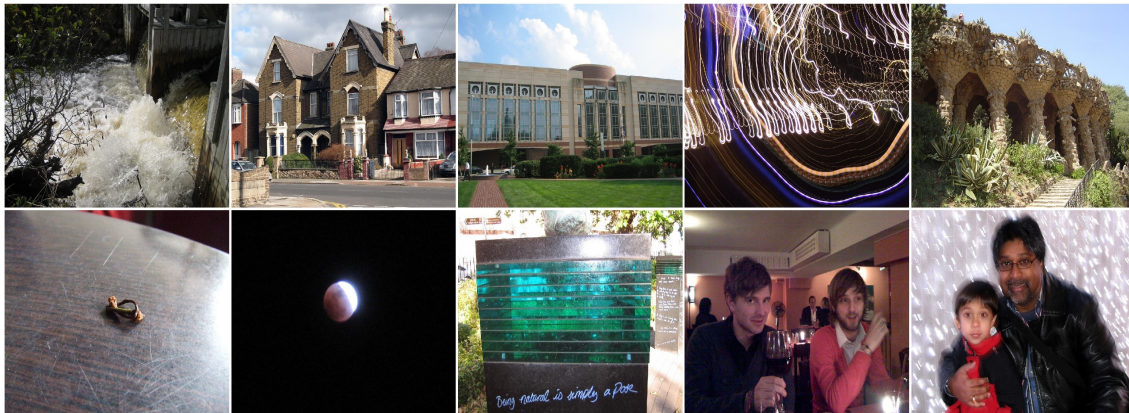
for the i th observer through ROC analysis. The AUC for the ROC curve is computed for all possible leave-one-out combinations. For AUC computation, we have used the shuffled AUC process employed in [4; 25]. The IOC of each image is given by averaging the AUC of all $i \in n$. In creating the predictive output from $(n-1)$ observers, the raw fixations of $(n-1)$ observers are convolved with a 2-D Gaussian distribution. The optimal size of Gaussian window was determined independently for each data set, in considering the maximal mean IOC score across the entire data set. The optimal σ varies in a fashion related to the resolution/scale of the input images. This provides a measure of the degree of agreement in viewing patterns across observers for each image within each dataset.

4.1.2 Experimental Methods and Results

To better understand the problem of IOC prediction, we have chosen three well-known and distinct datasets: The Bruce/Toronto dataset [8], the Mancas/LeMeur dataset [37] and the Judd/MIT dataset [24]. The Bruce/Toronto dataset contains 120 color images of 681×511 resolution with 20 human observers. Images from the Mancas and Le Meur [37] dataset include fixations for 17 observers, an image resolution of 384×384 pixels, and 135 total images. We have also considered 464 images (Landscape orientation) with resolution 1024×768 from the Judd/MIT dataset. The landscape orientation subset was selected to avoid diminished performance from including both landscape and portrait orientation. In particular given Gist features as one category for testing, the *squashing* of images to a square shape may present a disadvantage if differences in initial image sizes are present. Images across all databases include



(a)



(b)



(c)

Figure 4.1: Images are sorted based on their actual IOC score from (a) Bruce (b) Judd (c) Memorability datasets. Top rows corresponding to each dataset depict those images having low IOC and bottom rows depict images having high IOC. Specific properties common among higher or lower IOC scores at a semantic level are discussed in detail throughout this section.

a wide variety of scene categories and objects, and include natural scenes, indoor and outdoor scenes. There are some notable differences across the datasets, with one important difference being the relatively fewer instances of faces or text in the Bruce/Toronto dataset (This is evident in Fig. 4.1).

Single diagnostic features

As a starting point, we consider a number of simple features related to image complexity, some of which overlap with measures in the study of LeMeur et al. [32]. In this subsection, we describe these features and the underlying reasons for assuming a possible relation to IOC.

Entropy: Entropy is a measure of randomness, or local heterogeneity. Images typically exhibit correlation in their pixel values locally, and are far from being completely random. High entropy at the image level implies a weaker local correlation structure expressed broadly across the image. This will tend to occur in instances where there is a wide variety of objects/items/structure in the image, or sources of noise.

Visual clutter: Visual clutter is a quantitative measure of the level of ease to add a target that will draw attention within an existing image [42]. The clutter measure of Rosenholtz [42] is the most well established measure of visual clutter, and is therefore used to gauge the impact of image clutter on IOC. As clutter increases, one would naturally expect the associated IOC to fall given the distraction from clutter may imply less potential for individual items in the image to elicit perceptual contrast.

JPEG image size: The effect of image compression has been related to im-

age clutter [42] and is also intuitively related to scene complexity. Regularity in an image will tend to result in larger gains in compression. In this instance, we have employed the JPEG 2000 method due to its use of the Discrete Wavelet transform as a representation. Given a Wavelet decomposition as the basis for representation there is an inherent relationship to the spectral decomposition that one observes in early visual cortex, in the form of Gabor-like features with selectivity for angular and radial frequencies. An implication of this, is that one might expect the compressibility of an image as expressed by a spectral decomposition, carries diagnosticity at the level of the ability to have a relatively *silent* early cortical representation of a scene. For a fixed resolution, if the JPEG image size is high, one might expect that an image contains more irregular patterns, textures, or noise. In such a case, an additional factor of importance is the JPEG quality, or the number of coefficients that are driven to zero to boost compression performance at the expense of image quality. It is worth noting that this measurement resides at the level of simplistic pattern structure. More sophisticated mechanisms are likely at play in visual sampling by a human observer. Repeated patterns, and higher level cortical features may also minimize activity through recurrent disambiguation (e.g. as in the aperture problem), or through canonical operations such as divisive normalization [44; 41].

Without any learning these single diagnostic features demonstrate a significant degree of inverse correlation with IOC scores. It is somewhat surprising that some simple features are inherently tied to a complex prediction such as IOC. That said, a common denominator across each of these is the relation to sparsity, compressibility of signal, redundancy and efficient representation in an information theoretic sense.

There are numerous examples that suggest that these are some of the central tenets of cortical representation [20; 54; 39], which perhaps sheds light on the nature of this relationship. In Table 4.1, we present the Pearson correlation value (r) with confidence p_{val} for each single diagnostic features. We observe that JPEG image size and visual clutter have significant negative correlation with IOC. These values are also useful as a relative reference point for considering the performance of IOC prediction subject to richer feature sets that follow. We have also tested the correlation in applying varying degrees of JPEG compression. When quality is very low, many wavelet coefficients are ignored and below a certain threshold one may fail to adequately represent the image even at a relatively coarse grained level. When quality is increased above a certain threshold, then correlation does not change much with increasing quality. One does tend to observe a maximal degree of correlation for varying degrees of JPEG compression quality for any given dataset, that appears to be inversely related to image size. The experimental results are shown in Table 4.2. These results suggest a relative lack of importance of content beyond a certain high-frequency limit. This observation is consistent with the nature of foveation (that visual acuity is only high where one is currently looking), and that fixation patterns tend to be relatively stable as a function of image size up to a critical limit [22]. As a matter of comparison, features in the LeMeur et al. study [32] present r values for Pearson correlation of 0.34 for the Judd Dataset, and 0.27 for an alternative dataset [33]. It is also worth noting that this prior study used a single image set for training and evaluation, using the Bayesian Information Criterion to constrain model complexity.

It is interesting to note for these relatively simple features, that there is a stronger

Table 4.1: Correlation between actual IOC scores for single diagnostic features determined to have significant correlation with IOC.

Datasets	Bruce Dataset		Memorability Dataset		Judd Dataset	
Criteria	r	p_{val}	r	p_{val}	r	p_{val}
Entropy	-0.188114	0.039636	-0.082271	0.342813	-0.071841	0.122266
JPEG size	-0.327238	0.000264	-0.180868	0.035791	-0.196607	0.000002
Clutter FC	-0.34532	0.000112	-0.17086	0.047552	-0.187487	0.000048

Table 4.2: Correlation between actual IOC score and JPEG image size subject to varying degrees of JPEG quality

Datasets	Bruce Dataset		Memorability Dataset		Judd Dataset	
JPEG quality	r	p_{val}	r	p_{val}	r	p_{val}
1	-0.273497	0.002507	-0.119214	0.168458	-0.194706	0.000024
3	-0.292802	0.001173	-0.131575	0.128217	-0.20273	0.000011
5	-0.32086	0.000353	-0.16371	0.057791	-0.208112	0.000006
20	-0.340294	0.000143	-0.178525	0.038298	-0.200152	0.000014
45	-0.336347	0.000173	-0.178588	0.038228	-0.200425	0.000014
75	-0.327222	0.000264	-0.180971	0.035683	-0.196539	0.000000
100	-0.332018	0.000212	-0.187738	0.029223	-0.193499	0.000027

degree of correlation observed with the Bruce/Toronto dataset as compared with the 2 other datasets. It is conceivable that a weaker presence of *high-level* patterns (e.g. faces) is involved in this difference. In a free-viewing paradigm, one observes behavior that is distinct from explicit judgments of visual salience [29], and there is evidently the involvement of complex patterns involved in the range of behaviour observed.

More complex features

Having established that entropy, visual clutter and JPEG image size provide simple features with diagnostic value for predicting IOC, we now turn to the strategy of extracting richer feature vectors in combination with a learning algorithm to determine the value of scene level, structural, saliency based, or object based features in predicting IOC. The set of features used in predicting IOC through a learning

strategy follow:

Table 4.3: Correlation between actual IOC scores and predicted IOC scores using different *holistic* feature vectors for prediction.

Datasets	Bruce Dataset		Memorability Dataset		Judd Dataset	
Criteria	r	p_{val}	r	p_{val}	r	p_{val}
HoG	0.369516	0.003665	0.434057	0.000271	0.339512	0.299295
Gist	0.482696	0.000094	0.437934	0.000236	0.345628	0.015363
HoPS	0.504871	0.000039	0.470283	0.000068	0.430416	0.015592
DeepNet	0.330516	0.009902	0.370371	0.002205	0.397389	0.000000
DeepNet + HoPS	0.505926	0.000037	0.519035	0.000008	0.454831	0.000000

Bottom-Up and Top-Down Features

An important distinction that often appears within the visual attention literature, is the notion of bottom-up versus top-down processing. Bottom-up processing refers to the viewing behavior that is driven by the properties of the image content itself. A bright pattern, an unusual patch of color, or sudden movement tends to draw an observers gaze independent of any semantic information, or higher level concepts. Top-Down processing refers to the portion of viewing behavior that is driven by higher level cognition or prior knowledge. This may appear in the context of a certain task (e.g. looking for keys), or based on inherent reward seeking behaviour. For example, there is a strong tendency to look at faces or other socially relevant cues [32]. It is also observed that the overall holistic structure of a scene can be understood rapidly, and influence how a scene is examined [38]. To test the relative contribution of these different processes to observed viewing patterns, we therefore include a number of features that span these categories.

Saliency algorithms by design seek image content that draws interest from a stimulus driven perspective. Different models rely on different simple features, and mea-

asures of contrast with this common goal in mind. We therefore consider a rich bottom-up feature set derived from the combined output of a variety of popular saliency algorithms. The specifics of these features are described in more detail later in this section.

We also seek to include features that convey an overall structural representation of the scene. One popular method for representing the holistic scene envelope is through scene Gist [38]. This provides a set of holistic receptive fields that distinguishes between different categories of scene that carry different holistic structure. We have also used the more standard histogram of oriented gradients (HoG) [10] feature in a similar capacity. These representations are applied to the entire image to produce a feature vector that captures a coarse grained representation of scene structure.

For high-level features, we rely on object specific features that have shown significant success in large scale recognition tasks. These are features derived through deep learning using convolutional nets. High level features are derived from the BVLC Reference CaffeNet architecture [21] based on the AlexNet architecture [30] (and trained on the ILSVRC12 challenge data [43]). This provides a 1000 dimensional feature vector of responses corresponding to the highest layer of the deep convolutional net. The rationale for these features is to capture the presence of higher level concepts and patterns within the images, and to related these measurements to IOC.

Histogram of Predicted Saliency (HoPS) features:

IOC depends of variability in viewing patterns of different observers. Each observer may have their own way of analyzing a scene, that guides the viewing patterns observed in fixation data. As different saliency algorithms may emphasize different

types of feature contrast within an image, a feature level representation derived from a variety of saliency models is a natural option to consider. We have considered 12 different saliency algorithms in producing a feature vector to predict IOC. These algorithms include the Torralba [49], HouCVPR [16], HouNIPS [17], Itti-CIO2 [19], ImageSignatureLab [15], ImageSignatureRGB [15], SDSR [46], AIM [8], GBVS [14], AWS [13], Yan [56] and SOC [40] models. For a detailed description of these algorithm, the reader may refer to the original published work, or Borji et al. [4] for a summary. The HoPS features are computed as follows:

For each of the aforementioned saliency algorithms, a histogram is created that provides a summary representation of the distribution of predicted visual saliency within a rectangular region of the image. Histograms are normalized and subsequently concatenated to produce a feature set for prediction. As these features are derived from saliency maps, we call these features Histograms of Predicted Saliency (HoPS) in a manner similar to features based on histograms of normalized edge structure [10; 36]. Suppose, for the i th algorithm the normalized histogram is represented by $[f_1^i, f_2^i, f_3^i, \dots, f_b^i]$ where $i \in 1, 2, \dots, 12$ and b is number of bins in the histogram. The HoPS feature set is given by:

$$f = [f_1^1, f_2^1, \dots, f_b^1, f_1^2, f_2^2, \dots, f_b^2, \dots, f_1^{12}, f_2^{12}, \dots, f_b^{12}] \quad (4.1)$$

Where, any value f_j^i represents j th bin feature of the i th algorithm. If different algorithms agree with one other then this implies confidence across algorithms in a common region. Because of this characteristic, HoPS has very good capability to represent both variability in predicted saliency, but also consistency and commonality

across algorithms. Unlike the Gist features, HoPS features are computed without remapping to a fixed square resolution. This is important because resizing the saliency map to a square resolution can result in undue variability in scene level features in determining feature salience.

Learning: Extracted features are used to learn a regression model for IOC prediction. We have randomly selected half of the images from each dataset for training. Most of the images in any dataset exhibit a shuffled ROC score that falls within the range of 0.5 to 0.7. This may result in imbalance in the dataset in the relative representation of high and low IOC examples relative to average cases. To accommodate for bias in the distribution of IOC values, training is achieved in merging images of all three datasets. This results in a greater proportion of relatively high and low IOC values on a relative scale. As eye tracking data are derived from different numbers of observers for different datasets, we normalize the scale of IOC values via linear scaling for each dataset to a common range of 0 to 1. This helps to overcome the paucity of data inherent in performing an image-level prediction (as opposed to pixel level where there are many more samples). The limited size of datasets is the reason for a singular dataset in training and testing within the prior work of LeMeur et al [32], and this strategy allows us to successfully overcome this limitation. Repeated selection of sub-samples of the training set allows for the selection of training samples with a more uniform range following the same strategy employed by Isola et al. [18].

Given training instances represented by feature vectors (HoG, GIST, HoPS, Deep-Net) we apply ϵ -SVR [9] for continuous value prediction. The dimensionality of HoPS features may be altered by varying the number of bins. Performance is relatively con-

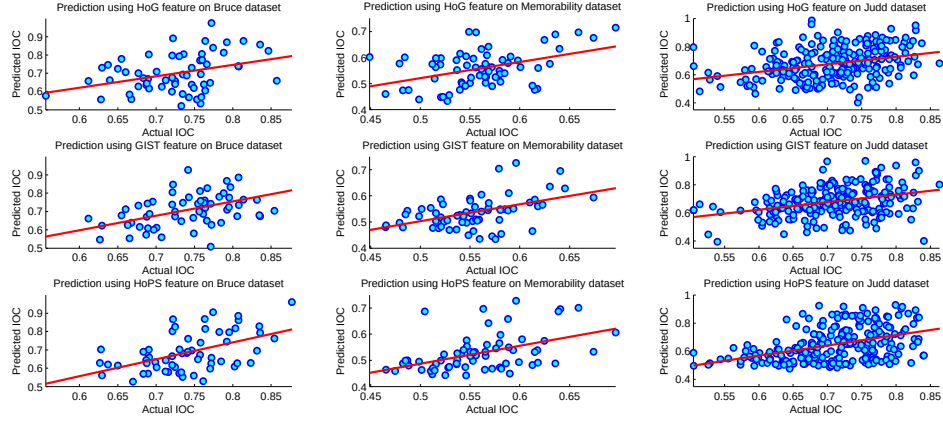


Figure 4.2: Predicted IOC vs. actual IOC for each combination of the three datasets, and HoG, GIST and HoPS features as predictors. The red line corresponds to the least squares linear regression fit to the data.

sistent across a range of bin values (5 to 100 bins were tested) with 20 bins used in test results. While either of HoPS or HoG may be applied subject to a grid or block-wise decomposition of the image, it was determined that there was no significant benefit to such a decomposition when compared with features of the same variety derived over the complete image. This is somewhat surprising, however the degree of contrast in the saliency maps is quite diagnostic of the presence of highly salient regions, independent of where they are located. Parameters were also optimized for each of HoG and Gist to achieve the greatest correlation with IOC scores. These results correspond to HoG using a cell size of 128 and 8 orientations. Similarly, the best performance of GIST for orientation per scale was $[2\ 2\ 2]$ (3 scales, 2 orientations per scale) and a subdivision using 4 blocks. As images from different datasets may have different resolution, these are resized to a fixed dimension for HoG and GIST which are 256×256 and 128×128 respectively. The results of these experiments are shown in Table 4.3. We observe that holistic structural feature extraction and sum-

mary statistics across saliency maps produce stronger correlation than that observed for the simpler features that were initially considered. Moreover, HoPS features outperform both HoG and GIST features suggesting that the distribution of saliency scores, and agreement across algorithms has strong diagnosticity for predicting IOC, and may also be of similar value for other perception based predictions. A scatter plot of predicted IOC vs. actual IOC value is depicted in Figure 4.2.

It is also interesting to note the contrast in the high level features derived from deep learning, and the HoPS features. While the HoPS features are of greatest value for the Bruce Dataset, they are of least value for the Judd dataset. The features derived from the deep neural network show the opposite trend, with these features having relatively less value for the Bruce dataset. This fits with the observation that more of the salient patterns in the Bruce dataset are a result of simpler feature configurations whereas the Judd dataset has a much higher incidence of people, faces and text. It is also noteworthy, that the best performance derives from combining the low-level stimulus driven features with the high-level object sensitive features. This underscores the importance of modeling both bottom-up and top-down facets of viewing behaviour.

In Figure 4.3, we show exemplar images that exhibit the greatest underestimate of calculated IOC based on the residuals from the regression fit. These are selected from the Judd dataset, owing to the greater number of images, and variety of content in this data with respect to semantically meaningful patterns that seem to drive fixation bias. The 15 images corresponding to the most underestimated IOC values were selected for each of the HoPS and deep learning based predictions. We show

the 7 largest differences in common to both the HoPS and deep learning features (a), those that appear only in the bottom 15 for the HoPS based predictions (b), and those that appear uniquely in the bottom 15 for the deep learning based predictions (c). The largest underestimates specific to HoPS are dominated by images containing text faces. Underestimates for the deep learning features are a bit more varied. One of these corresponds to a simple pattern that is highly salient based on bottom-up factors. Others contain instances of objects that may be readily interpreted by a human observer, but may be small in the context of the overall scene, or relatively low in contrast. There are also instances of faces, but it is notable that these represent a blurry photo, atypical pose, and a dog face with partial occlusion of the face by hair. Underestimates common to both types of features seem to correspond to scenes with semantically meaningful content that are relatively complex in their structure or lighting conditions. There is some suggestions from these cases that both bottom-up and top-down features are important to predictive modeling of gaze behavior based on image content.

4.1.3 Discussion

In this section, we have explored the problem of predicting IOC among human observers based on previously unseen images. The extent to which this characterization may be made evidently depends in part on the nature of the images under consideration. IOC correlation associated with a number of low-level simplistic features hints at the importance of redundancy and neural coding in how a scene is parsed by a human observer. The success of saliency models built on principles from informa-

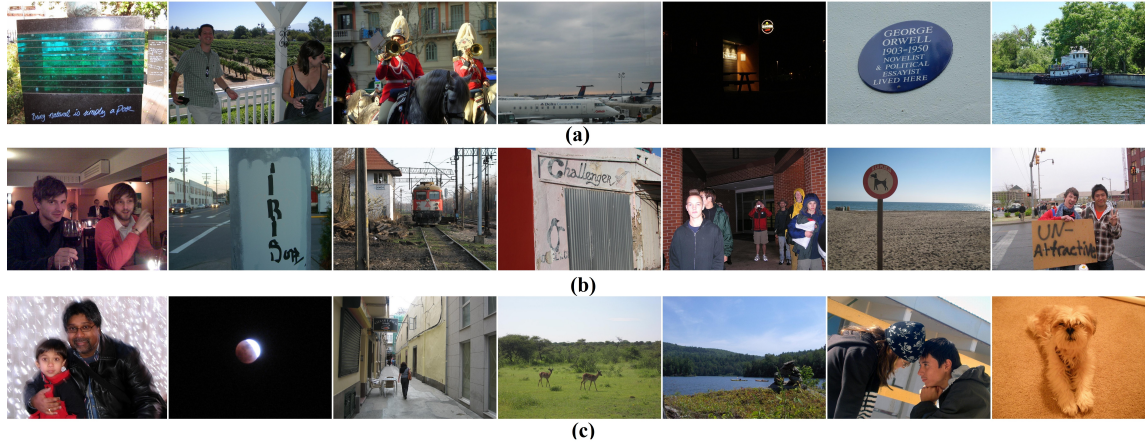


Figure 4.3: Images among the greatest 15 underestimates of actual IOC scores (based on residuals from the regression model): (a) Common to both HoPS and Deep learning features (b) Within the bottom 15 for HoPS only. (c) Within the bottom 15 for Deep learning only.

tion theory, compressive sensing and coding, (divisive) normalization and whitening evidently mirrors some of the characteristics that are relevant in characterizing gaze behavior in a holistic fashion.

We have also introduced a novel feature representation characterized by the distribution of values produced by different saliency models for a particular image. This provides a feature with strong diagnostic value in predicting IOC. It is also reasonable to assume that HoPS type features may have value in alternative predictions related to perception.

The value of features in regression, and dependence on the type of data carries important implications for future work in computational models of perception. While local contrast (or saliency) measures carry significant value in predictive models, there are evidently many other elements beyond context or scene structure that factor into viewing behaviour. There are a multitude of inherent behaviours relating to reward

seeking, fear or social interaction at play even within a *free viewing* context. The value of more semantically grounded features drawn from a deep learning paradigm underscores the value in marrying low-level image characteristics with higher level image understanding. This is important in the prediction of IOC, but also more generally in systems for active vision, early proto-object selection and any models targeting perceptual computation.

4.2 Other subjective ratings

In the era of the internet, we now have access to millions of images. Given the large number of images people are now exposed to, this necessitates paying attention to any given image for a very short amount of time. However, people may experience an affective response to any image subject to a very short time course. These subjective feelings can be motivated by personal experience, social factors and the form and content of the image. The ability to predict the subjective experience of people to images without the need for a time consuming user survey is a valuable tool for managing digital media. In some recent work, specific instances of this problem are considered that focus on the *popularity* and *memorability* of images respectively [28; 18].

In the same vein, we address a variety of other subjective criteria including brand experience, purchase intention, interest, understanding, valence, relevance and originality. Images consist of 151 magazine pages from Swedish magazines, and locally distributed flyers for product advertising. In each case, participants were asked to provide subjective ratings on the aforementioned categories on different numeric scales.

Generally this scale consisted of a 1-5 rating, however the valence rating was based on a -5 to +5 scale for better intuitive correspondence to positive and negative valence. Memorability scores are more objective and reflect the capacity to discriminate between examples appearing in an experimental trial done at an earlier time. In short, the characteristics that these ratings seek to measure are as follows:

- Brand experience: To what extent does the participant have prior experience or familiarity with the products featured in the image
- Purchase intention: Does the participant have any intent to purchase the product in the future?
- Interest: What level of interest does the form/content of the image elicit?
- Understanding: Is the message of the advertisement clear to the participant? Are the product advertised, and salient features clear?
- Valence: How does the participant feel in viewing the image? (e.g. positive feelings, negative feelings)
- Relevance: How relevant is the content of the advertisement to the product that is featured?
- Originality: Is the advertisement original or does it stand out from other advertisements as unique, or creative?

4.2.1 Prediction model

Feature extraction supporting different types of subject rating prediction are based on the same features considered for the IOC case, as described in Chapter 4.1.2. As done for IOC, we consider HoG, GIST, HoPS and deepNet features for learning. This includes extraction of those features from the images themselves, but also based on saliency maps produced by different algorithms, to characterize holistic structure of the saliency output. The HoPS like feature extraction from color images is equivalent to extracting histograms from different subdivisions of an image when applied to the image rather than the saliency map, and is referred to as image histogram in the list of features. For learning, we have used half of the images for training and other half for testing. With the different types of extracted features, we have trained an ensemble of bagged decision trees for regression (Random Forest [6]). We have performed separate training and testing for each of the subjective rating types.

4.2.2 t-stochastic neighbor embedding (tSNE) of image features

We have extracted features from both image and saliency maps of the algorithms, and these represent the image contents from low level to high level features, with different relations to the subject rating assigned by people. Based on the properties each type of feature highlights different characteristics of the nature of images that may express similarity in structure, edge content, colors, or objects present. To visualize the relatedness of images subject to different types of feature representation, we embed all testing images within a 2D plot. The embeddings are done by t-

stochastic neighbor embedding (tSNE) [53]. The embedding for deepNet, raw image Gist, saliency map Gist, raw image HoG, saliency map HoG, raw image histogram and saliency map HoPS feature are shown in Figures 4.4, 4.5, 4.6, 4.7, 4.8, 4.9 and 4.10 respectively.

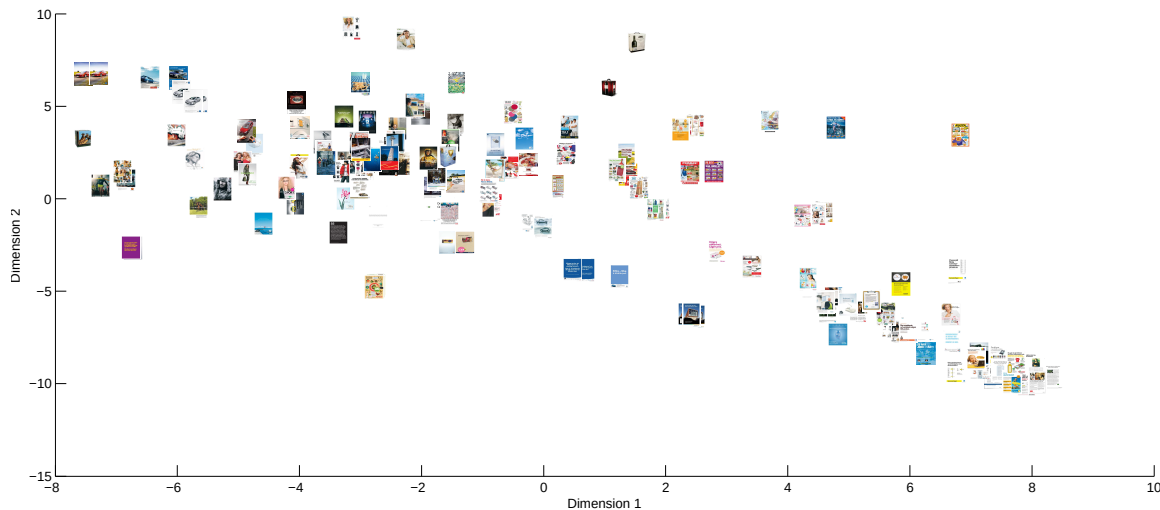


Figure 4.4: 2D Embedding of deepNet feature by tSNE. DeepNet feature captures mostly the high level features. Because of this dominant objects of similar nature (like human, car) are grouped together in the embedding

4.2.3 Experimental results

As stated, the parameter settings for different types of feature extraction mirrors the description appearing in Chapter 4.1.2. For features pertaining to learning, we have used 200 trees in the random forest. In Table 4.4, we present the best results after tuning parameters. We have only presented those results for which the correlation achieves more than 95% confidence. We notice that for some subject rating criteria like interest, understanding, memory and purchase intention, we are unable to achieve good correlation performance (corresponding to blank space in Table 4.4).

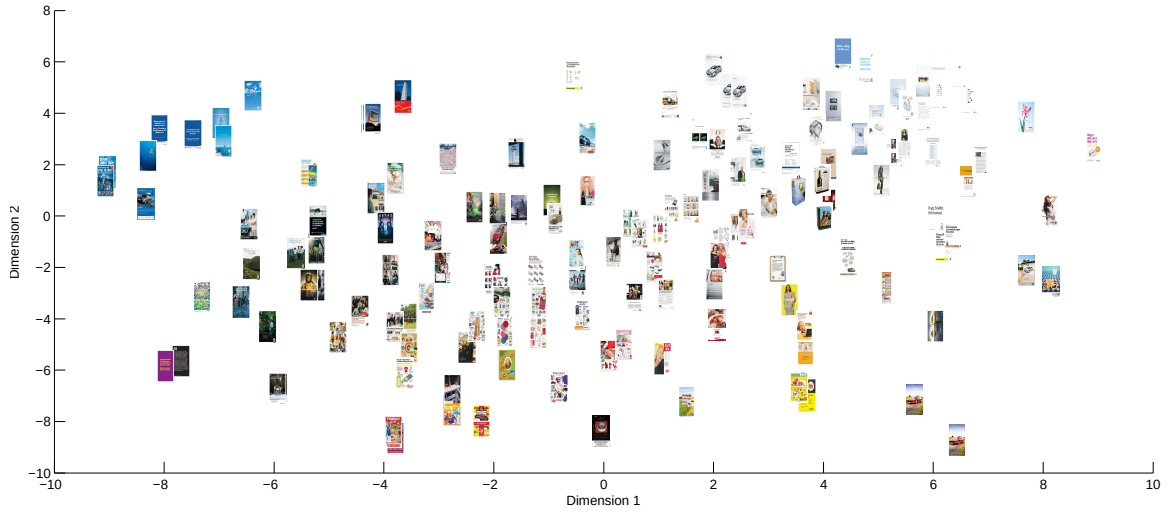


Figure 4.5: 2D Embedding of image Gist feature by tSNE. It is quite noticeable that global color directs the embedding. As Gist is applied on raw image, color becomes the distinguishing factor

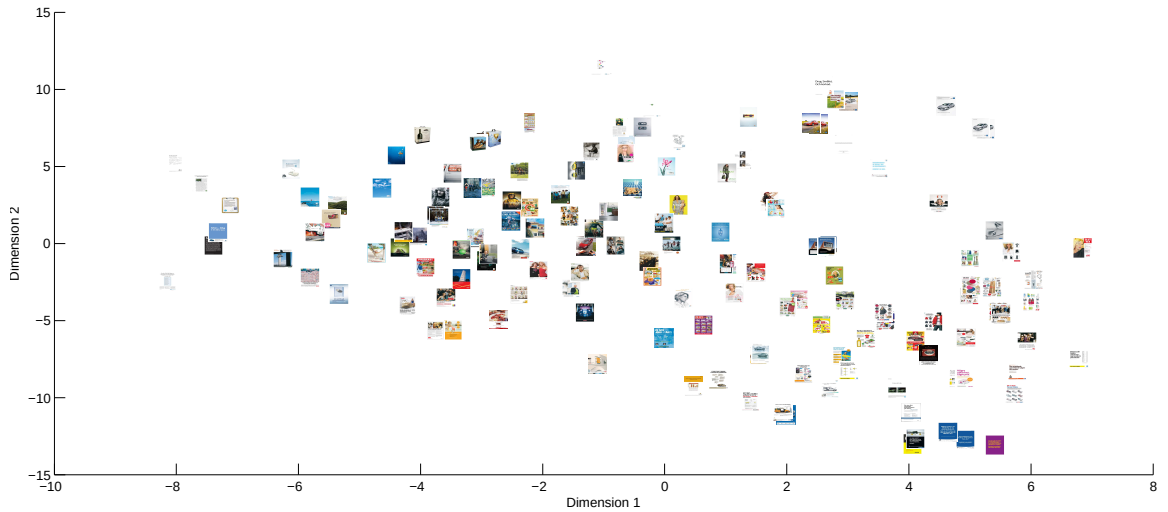


Figure 4.6: 2D Embedding of saliency Gist feature by tSNE. Images having similar looking saliency map are embedded in close locations

The low correlation for these factors indicates that for the magazine advertisements, those subjective ratings may be more strongly dominated by personal experience, social cues, individual choice and other top-down factors which are not discernable from the image content alone. On the other hand, some subjective criteria like brand

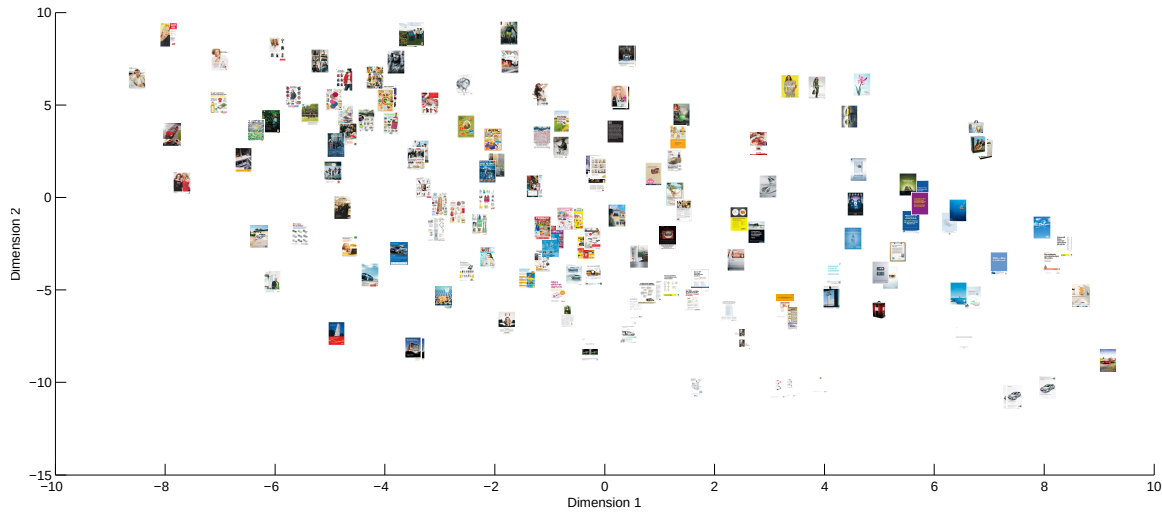


Figure 4.7: 2D Embedding of image HoG feature by tSNE. Images having similar local edge structure are in close location

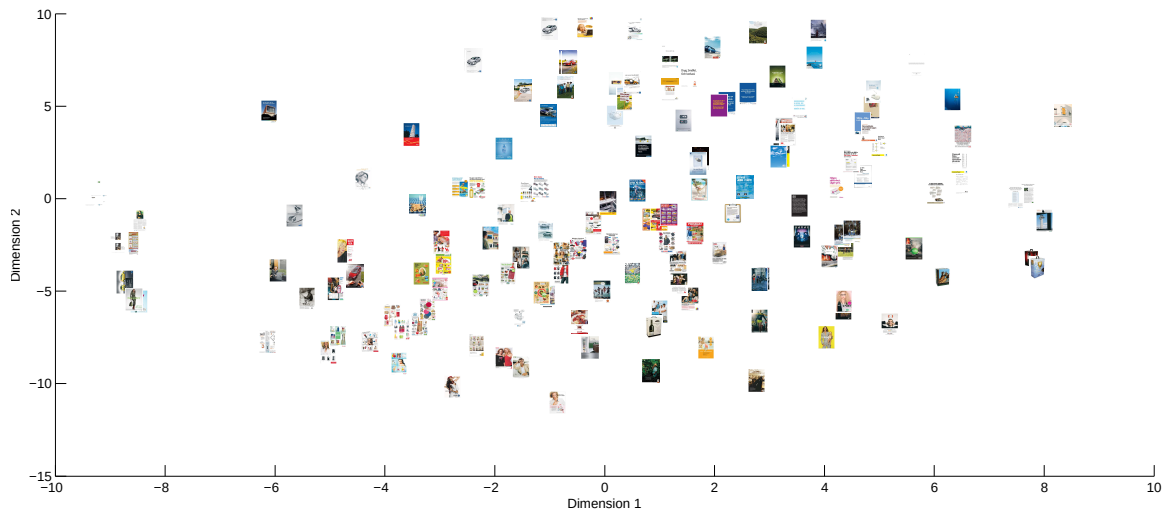


Figure 4.8: 2D Embedding of saliency HoG feature by tSNE. Images having similar looking edges in saliency map are embedded in close locations

experience, relevance, valence and originality achieve relatively consistent correlation (with confidence) for different feature types. This indicates that the image itself contains significant information concerning factors that drive these subjective ratings. In particular, brand experience and originality are noticed to be highly predictable.

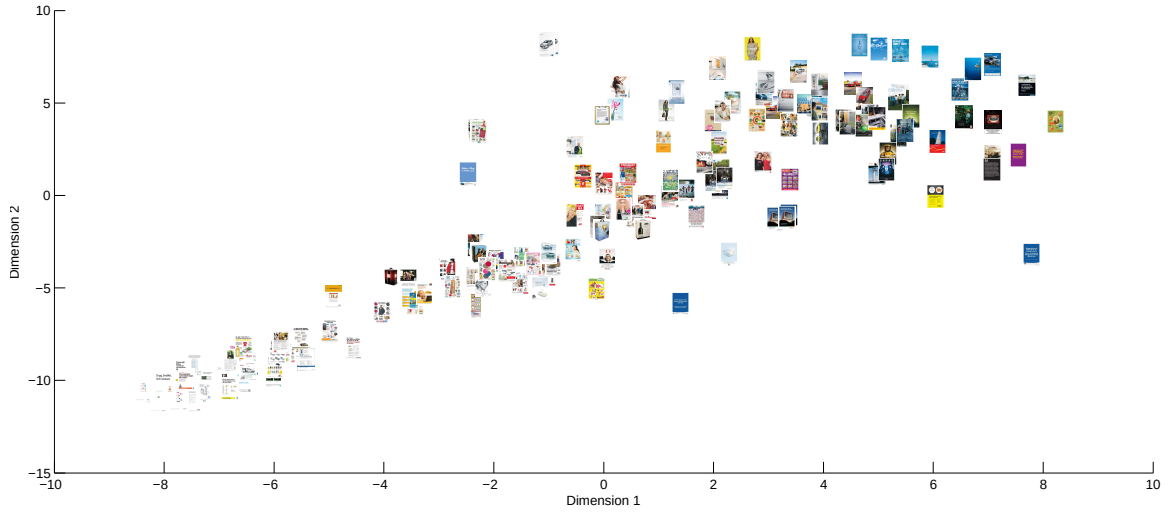


Figure 4.9: 2D Embedding of image histogram feature by tSNE. This embedding is dominated by both color and local structure

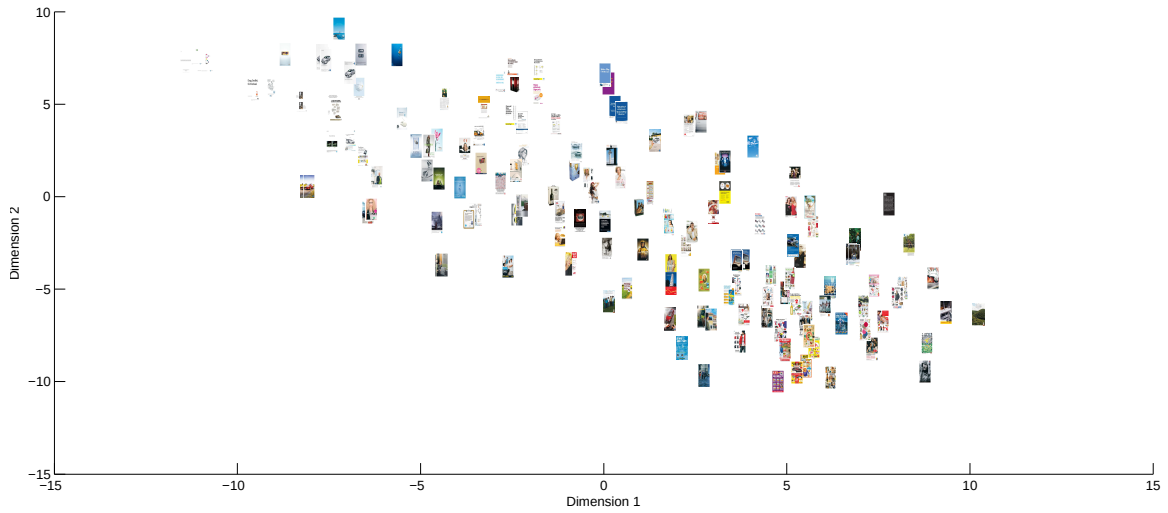


Figure 4.10: 2D Embedding of saliency HoPS feature by tSNE. Images having similar looking saliency map are embedded in close locations

Intuitively, originality may have some relation to the spatial organization of content in the image. Successful prediction of brand experience is more surprising. However, in examining the images themselves, well known brands seem to preferentially employ advertisements with a relatively clean appearance, and highly visible logo and

message. For this reason, predictability is not necessarily related to direct causality, but rather to correlation due to bias in how advertisements are composed according to the brands they represent.

Table 4.4: Prediction performance as Pearson Correlation of subjective rating of different types of features

Rating Type	Image Gist	Saliency Gist	Image HoG	Saliency HoG	Image Hist	Saliency HoPS	Deep Net
Brand Experience	0.35	0.41	-	0.30	0.34	0.39	0.32
Interest	-	-	-	-	-	0.27	0.28
Understanding	0.29	-	-	-	-	-	-
Relevance	0.33	0.42	0.32	0.40	-	0.38	-
Valence	0.28	0.28	-	-	-	-	0.37
Memory	-	-	-	-	-	-	0.29
Purchase Intention	-	0.28	0.26	-	-	-	-
Originality	0.43	0.52	0.33	0.44	0.45	0.52	0.53

Theoretically, we know that concatenation of different useful features can improve the performance of a prediction model. To test this, we also experiment with different combinations of high performing feature types to improve on individual performance of the features. We have done this experiment for the subjective rating types which are consistently predictable by individual features. In Table 4.5, 4.6, 4.7, 4.8, we report feature combination results of brand experience, relevance, valence and originality of image. We notice that in most of the cases the combination of features improves on correlation relative to the correlation of individual features.

4.2.4 Discussion

In this section, we have addressed different subjective ratings associated with images and determine the value of different features in successfully approximating these ratings. We use a wide range of subjective rating criteria including brand experi-

Table 4.5: Testing with combination of features for rating type Brand Experience

Combination	Feature Type	Pearson Correlation
1st best	Sal Gist	0.41
2nd best	Sal HoPS	0.39
3rd best	Im Gist	0.35
1st+2nd best	Sal Gist+Sal HoPS	0.44
1st+3rd best	Sal Gist+Im Gist	0.35
2nd+3rd best	Sal HoPS+Im Gist	0.36
1st+2nd+3rd best	Sal Gist+Sal HoPS+Im Gist	0.41

Table 4.6: Testing with combination of features for rating type Relevance

Combination	Feature Type	Correlation
1st best	Sal Gist	0.42
2nd best	Sal HoG	0.40
3rd best	Sal HoPS	0.38
1st+2nd best	Sal Gist+Sal HoG	0.39
1st+3rd best	Sal Gist+Sal HoPS	0.37
2nd+3rd best	Sal HoG+Sal HoPS	0.42
1st+2nd+3rd best	Sal Gist+Sal HoG+Sal HoPS	0.42

ence, purchase intention, interest, understanding, valance, relevance and originality. Among them, a handful of criteria including brand experience, valence, relevance and originality are consistently predictable with features that include Gist, HoG and HoPS derived from both images and saliency maps. Moreover, concatenation of different

Table 4.7: Testing with combination of features for rating type Valence

Combination	Feature Type	Pearson Correlation
1st best	DeepNet	0.37
2nd best	Im Gist	0.28
3rd best	Sal Gist	0.28
1st+2nd best	DeepNet+Im Gist	0.37
1st+3rd best	DeepNet+Sal Gist	0.39
2nd+3rd best	Im Gist+Sal Gist	0.29
1st+2nd+3rd best	DeepNet+Im Gist+Sal Gist	0.41

Table 4.8: Testing with combination of features for rating type Originality

Combination	Feature Type	Pearson Correlation
1st best	DeepNet	0.53
2nd best	Sal HoPS	0.52
3rd best	Sal Gist	0.52
1st+2nd best	DeepNet+Sal HoPS	0.59
1st+3rd best	DeepNet+Sal Gist	0.58
2nd+3rd best	Sal HoPS+Sal Gist	0.54
1st+2nd+3rd best	DeepNet+Sal HoPS+Sal Gist	0.60

combinations of features helps to improve prediction performance. Prediction of image level subjective criteria is very valuable to the advertising industry because based on the capacity to predict effectiveness a priori, and before committing expenses to circulating the advertisement. The value of different types of features also sheds light on effective strategies for sentiment based prediction in general.

Chapter 5

Summary, Limitations and Future Work

5.1 Summary

As a whole, this research enriches our understanding of visual saliency computation and knowledge concerning the perception of digital media. This includes a better understanding of reasons for success and failure of different established algorithms. We demonstrate our hypotheses in considering a range of imagery that includes both natural and man-made content offering deeper insight into the landscape of how different algorithms behave. In addition, we attempt to improve the performance of human-like saliency computation, in better demonstrating the importance of divisive normalization and second order contrast measures [27]. Relevant to general standards for performance evaluation in this area, confounds imposed by central bias are solved through novel strategies to gauge the spatial bias produced by algorithms

and remove center bias in evaluation. Moreover, in regards to predictive methods targeting digital media, we present novel models for predicting explicit judgment locations, IOC and other subjective ratings at the *whole image* level. These predictive models are validated through comparison with human data from user studies. This thesis generates a number of new directions in image saliency research, and presents some additional baseline approaches targeting new types of predictive analysis for characterizing perceptually relevant image properties.

5.2 Limitations

As with virtually any research endeavor, there are additional avenues that would benefit from further analysis. To this end the following identifies a few specific targets for further investigation associated with the content presented in this thesis.

- We tune the parameters of the SOC model based on DOVES images, which is a dataset of *natural images*. However, testing with Toronto dataset uses parameters tuned to natural images. With further exploration of the parameter space, the performance of the SOC model would have been better for the case of Toronto dataset.
- While learning prediction models we have not calculated the feature importance during training. It might help us know which specific features are most discriminative, to propose a reduced feature set for different prediction criteria.
- In this thesis, while training any prediction model, we have applied less rigour to the learning strategy, with the focus on features and motivating ideas. We

have used linear SVR, and standard random forest implementation. Learning performance may improve by defining the formulation our problems in support of structured prediction, transfer learning and learning with hidden variables like latent SVM. However, the methods used are sufficient to draw some valuable conclusions concerning different aspects of perception of digital media in what has been considered.

5.3 Future Work

There exist many possibilities for future research derived from this thesis, and a set of these possibilities are listed below:

- Divisive normalization and second order contrast can improve correspondence in the match between image level outputs, and measured neural responses for natural stimuli. We find that such improvement is also reflected in saliency detection. Similarly, the difference in including local divisive normalization, and second order contrast in particular may be of value to alternative problems such as identifying natural images, or distinguishing between indoor and outdoor settings. In this way, we can develop biologically motivated natural image classification system and other relevant applications.
- We show that removal center bias significantly affects the ranking of algorithms of saliency detection. In prior literature, saliency algorithms are also used in object segmentation [34]. The role of center bias in salient object detection has not been examined in the same degree of detail, and performance of object

segmentation algorithms might be assessed with similar treatment of center bias.

- Saliency models tend to exhibit a high degree of sensitivity to imaging conditions [3]. Keeping this in mind, an user study based on different imaging conditions for different perceptual task like free viewing, object searching, saliency viewing and explicit judgment would be of value to gauge what this variation looks like for human observers. This may also reveal how saliency algorithm should treat affine transformations, or lighting variations with sensitivity to either performance as a component of a larger vision system, or in fidelity to human behaviour.
- Predictions of subjective ratings of digital media can be improved by learning deep neural network models for this purpose. The convolution layers of deep networks can reveal bottom-up low level features and top layers can reveal top-down high level features which are evidently both important from the results of experiments included in this thesis. The primary limitation at this time in considering such an approach is the unavailability of a dataset that includes a sufficient number of samples to allow for this type of training.

Bibliography

- [1] ACHANTAY, R., HEMAMIZ, S., ESTRADAY, F., AND SSSTRUNKY, S. Frequency-tuned salient region detection. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009* (2009), 1597–1604.
- [2] ALPERT, S., GALUN, M., BASRI, R., AND BRANDT, A. Image segmentation by probabilistic bottom-up aggregation and cue integration. 1–8.
- [3] ANDREOPOULOS, A., AND TSOTSOS, J. K. On sensor bias in experimental methods for comparing interest-point, saliency, and recognition algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34, 1 (2012), 110–126.
- [4] BORJI, A., SIHITE, D., AND ITTI, L. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing* 22, 1 (2013), 55–69.
- [5] BORJI, A., SIHITE, D. N., AND ITTI, L. What stands out in a scene? a study of human explicit saliency judgment. *Vision Research* 91, 0 (2013), 62 – 77.

-
- [6] BREIMAN, L. Random forests. *Machine learning* 45, 1 (2001), 5–32.
 - [7] BRUCE, N. D., WLOKA, C., FROSST, N., RAHMAN, S., AND TSOTSOS, J. K. On computational modeling of visual saliency: Examining whats right, and whats left. *Vision Research*, 0 (2015), –.
 - [8] BRUCE, N. D. B., AND TSOTSOS, J. K. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision* 9, 3 (2009).
 - [9] 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.
 - [10] FELZENSZWALB, P. F., GIRSHICK, R. B., MCALLESTER, D., AND RAMANAN, D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 9 (2010), 1627–1645.
 - [11] FIELD, D. J. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America. A, Optics and image science* 4, 12 (1987), 2379–2394.
 - [12] GARCIA-DIAZ, A., FDEZ-VIDAL, X., PARDO, X., AND DOSIL, R. Decorrelation and distinctiveness provide with human-like saliency. *Lecture Notes in Computer Science 5807 LNCS* (2009), 343–354.
 - [13] GARCIA-DIAZ, A., FDEZ-VIDAL, X., PARDO, X., AND DOSIL, R. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing* 30, 1 (2012), 51–64.

-
- [14] HAREL, J., KOCH, C., AND PERONA, P. Graph-based visual saliency. *Advances in neural information processing systems 19* (2007), 545.
 - [15] HOU, X., HAREL, J., AND KOCH, C. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 1 (2012), 194–201.
 - [16] HOU, X., AND ZHANG, L. Saliency detection: A spectral residual approach. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2007).
 - [17] HOU, X., AND ZHANG, L. Dynamic visual attention: Searching for coding length increments. pp. 681–688.
 - [18] ISOLA, P., XIAO, J., TORRALBA, A., AND OLIVA, A. What makes an image memorable? *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2011), 145–152.
 - [19] 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.
 - [20] JAZAYERI, M., AND MOVSHON, J. A. Optimal representation of sensory information by neural populations. *Nature Neuroscience* 9, 5 (2006), 690–696.
 - [21] JIA, Y., SHELHAMER, E., DONAHUE, J., KARAYEV, S., LONG, J., GIRSHICK, R., GUADARRAMA, S., AND DARRELL, T. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093* (2014).

-
- [22] JUDD, T., DURAND, F., AND TORRALBA, A. Fixations on low-resolution images. *Journal of Vision* 11, 4 (2011), 1–20.
 - [23] JUDD, T., DURAND, F., AND TORRALBA, A. A benchmark of computational models of saliency to predict human fixations. *MIT Computer Science and Artificial Intelligence Lab Technical Report* (2012).
 - [24] JUDD, T., EHINGER, K., DURAND, F., AND TORRALBA, A. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on* (2009), IEEE, pp. 2106–2113.
 - [25] KANAN, C., TONG, M., ZHANG, L., AND COTTRELL, G. Sun: Top-down saliency using natural statistics. *Visual Cognition* 17, 6-7 (2009), 979–1003.
 - [26] KANAN, C., TONG, M. H., ZHANG, L., AND COTTRELL, G. W. Sun: Top-down saliency using natural statistics. *Visual Cognition* 17, 6-7 (2009), 979–1003.
 - [27] KAY, K. N., WINAWER, J., ROKEM, A., MEZER, A., AND WANDELL, B. A. A two-stage cascade model of bold responses in human visual cortex. *PLoS Computational Biology* 9, 5 (2013).
 - [28] KHOSLA, A., DAS SARMA, A., AND HAMID, R. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web* (2014), International World Wide Web Conferences Steering Committee, pp. 867–876.
 - [29] KOEHLER, K., GUO, F., ZHANG, S., AND ECKSTEIN, M. P. What do saliency models predict? *Journal of Vision* 14, 3 (2014).

-
- [30] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105.
- [31] LE MEUR, O., AND BACCINO, T. Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods* 45, 1 (2013), 251–266.
- [32] LE MEUR, O., BACCINO, T., AND ROUMY, A. Prediction of the inter-observer visual congruency (iovc) and application to image ranking. In *Proceedings of the 19th ACM international conference on Multimedia* (2011), ACM, pp. 373–382.
- [33] LE MEUR, O., LE CALLET, P., BARBA, D., AND THOREAU, D. A coherent computational approach to model bottom-up visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 28, 5 (2006), 802–817.
- [34] LI, Y., HOU, X., KOCH, C., REHG, J., AND YUILLE, A. The secrets of salient object segmentation. pp. 280–287.
- [35] 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.
- [36] LOWE, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* 60, 2 (Nov. 2004), 91–110.

-
- [37] MANCAS, M., AND LE MEUR, O. Memorability of natural scenes: The role of attention. *2013 IEEE International Conference on Image Processing, ICIP 2013 - Proceedings* (2013), 196–200.
- [38] OLIVA, A., AND TORRALBA, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42, 3 (2001), 145–175.
- [39] OLSHAUSEN, B. A., AND FIELD, D. J. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research* 37, 23 (1997), 3311 – 3325.
- [40] RAHMAN, S., ROCHAN, M., WANG, Y., AND BRUCE, N. Examining visual saliency prediction in naturalistic scenes. In *Image Processing (ICIP), 2014 IEEE International Conference on* (Oct 2014), pp. 4082–4086.
- [41] REYNOLDS, J. H., AND HEEGER, D. J. The normalization model of attention. *Neuron* 61, 2 (2009), 168 – 185.
- [42] ROSENHOLTZ, R., LI, Y., AND NAKANO, L. Measuring visual clutter. *Journal of Vision* 7, 2 (2007).
- [43] 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, 2014.
- [44] SCHWARTZ, O., AND SIMONCELLI, E. P. Natural signal statistics and sensory gain control. *Nature neuroscience* 4, 8 (2001), 819–825.

-
- [45] SEO, H., AND MILANFAR, P. Nonparametric bottom-up saliency detection by self-resemblance. *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009* (2009), 45–52.
- [46] SEO, H. J., AND MILANFAR, P. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision* 9, 12 (2009), 1–27.
- [47] TATLER, B. W. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision* 7, 14 (2007).
- [48] TATLER, B. W., BADDELEY, R. J., AND GILCHRIST, I. D. Visual correlates of fixation selection: effects of scale and time. *Vision Research* 45, 5 (2005), 643 – 659.
- [49] TORRALBA, A. Modeling global scene factors in attention. *Journal of the Optical Society of America A: Optics and Image Science, and Vision* 20, 7 (2003), 1407–1418.
- [50] TORRALBA, A., AND OLIVA, A. Statistics of natural image categories. *Network: computation in neural systems* 14, 3 (2003), 391–412.
- [51] 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.
- [52] VAN DER LINDE, I., RAJASHEKAR, U., BOVIK, A. C., AND CORMACK, L. K. Doves: A database of visual eye movements. *Spatial Vision* 22, 2 (2009), 161–177.

-
- [53] VAN DER MAATEN, L., AND HINTON, G. Visualizing data using t-sne. *Journal of Machine Learning Research* 9, 2579-2605 (2008), 85.
- [54] WAINWRIGHT, M. J. Visual adaptation as optimal information transmission. *Vision Research* 39, 23 (1999), 3960–3974.
- [55] WANG, Z., AND LI, B. A two-stage approach to saliency detection in images. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (2008), 965–968.
- [56] YAN, J., LIU, J., LI, Y., NIU, Z., AND LIU, Y. Visual saliency detection via rank-sparsity decomposition. *Proceedings of International Conference on Image Processing, ICIP* (2010), 1089–1092.
- [57] ZHANG, L., TONG, M., MARKS, T., SHAN, H., AND COTTRELL, G. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision* 8, 7 (2008).
- [58] ZHAO, Q., AND KOCH, C. Learning a saliency map using fixated locations in natural scenes. *Journal of vision* 11, 3 (2011).