The Eyes as a Window to the Mind: Inferring Cognitive State from Gaze Patterns

by

Jonathan F.G. Boisvert

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
December 2015

Thesis advisor

Dr. Neil Bruce

Author

Jonathan F.G. Boisvert

The Eyes as a Window to the Mind:
Inferring Cognitive State from Gaze Patterns

Abstract

In seminal work, Yarbus examined the characteristic scanpaths that result when viewing an image, observing that scanpaths varied significantly depending on the question posed to the observer. While early efforts examining this hypothesis were equivocal, it has since been established that aspects of an observer's assigned task may be inferred from their gaze. In this thesis we examine two datasets that have not previously been considered involving prediction of task and observer sentiment respectively. The first of these involves predicting general tasks assigned to observers viewing images, and the other predicting subjective ratings recorded after viewing advertisements. The results present interesting observations on task groupings and affective dimensions of images, and the value of various measurements (gaze or image based) in making these predictions. Analysis also demonstrates the importance of how data is partitioned for predictive analysis, and the complementary nature of gaze specific and image derived features.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to begin by thanking my advisor, my committee, my parents, and all the people who have supported me along the way.

This thesis is dedicated to somebody special. You know who you are.

Chapter 1

Introduction

Early seminal work in analyzing eye movement patterns by Buswell [1935] and Yarbus [1967] remains influential in shaping scientific discourse addressing the role of task in gaze behaviour. This includes the notion that an observer's task may be inferred from examining their eye movements. Results from Yarbus' early work, depicted in Fig. 1.1, suggest a probable link between cognitive processes or intent, and observed eye movements.

Many research efforts have further examined the interaction between task and fixations, in some cases directly considering Yarbus' claims about the predictability of task from fixations [Castelhano et al., 2009; Tatler et al., 2010]. Some of this analysis has leveraged modern techniques in pattern classification to directly predict task from recorded fixation data. One prior effort modelled heavily on the methodology of Yarbus' experiments, [Greene et al., 2012], considers three different classifiers applied to aggregated eye-movement statistics for task prediction. None of these classifiers yielded performance above chance in considering the aggregate fixation features, calling into question the validity of Yarbus' assertions on task predictability from gaze. More recent work considering the same data, but instead using

low-resolution fixation density patterns [Borji and Itti, 2014], or other statistics achieved above chance performance by revising the feature set and inference methods used. Further improvements for the same data set have also been achieved in assuming knowledge of the participant viewing the image, or the specific image under consideration [Kanan et al., 2014]. Attempts to include a representation of covert attention in models of this variety has also lead to higher prediction accuracies [Haji-Abolhassani and Clark, 2014]. The data provided by Greene et al. [2012] has also inspired analysis of the possible impact of tasks requiring both visual processing and language processing [Coco and Keller, 2014]. There are evidently important individual differences that factor into viewing patterns [Tatler et al., 2005], and such results also point to the notion that features such as fixation position and duration carry significant information about a viewer's assigned task.

In the context of attention modeling, there are many studies that address the relative contribution of *bottom-up* versus *top-down* influences on the deployment of overt attention and fixation patterns. The spirit of Yarbus' assertions emphasis the role of an active overt attention process, drawing heavy influence from top-down bias. The critical importance of task in attention, and overt attention specifically is well supported as discussed in a number of recent studies and review papers addressing the relative importance of task and top-down cues. [Hayhoe and Ballard, 2005; Chen and Zelinsky, 2006; Rothkopf et al., 2007; Tatler and Vincent, 2009; Yang and Zelinsky, 2009; Tatler et al., 2011].

While recent efforts leave little doubt that observed fixations provide a window into cognitive state or task directives, there is remaining benefit in examining task predictability from gaze statistics. One evident benefit from an applied perspective is the the capability to infer task or intentions from fixations for applications in human machine interaction and human centric computing. Additional benefits of a more general nature arise from examining

the ease with which different tasks may be distinguished based on gaze patterns, and in determining which features successfully discriminate between tasks. The degree of task separability has value in understanding similarity among visual and attentive mechanisms recruited for different tasks. Determining specific factors that distinguish tasks also points to targets for more careful examination in targeted experimental studies involving humans (e.g. psychophysics). In the domain of task prediction, prior work has focused heavily on confirming or denying the hypothesis that task may be predicted from gaze. For this reason, there has been a strong emphasis on prediction accuracy with less consideration of the role of different image or gaze related statistics in determining prediction performance. The work presented in this thesis expands on the body of research involving task prediction in addressing the following important questions:

1. **Do relatively coarse grained tasks present distinct gaze statistics?** While several different sets of tasks have been examined in the literature, we consider task definitions that reside at a relatively general or coarse-grained level of abstraction. This serves to contribute to the growing body of efforts examining task prediction, and also to add diversity in the types of task sets considered. Some more fine-grained task definitions are also sufficiently distinct that it would be surprising if very different gaze patterns were not observed.

2. **What methodological considerations are most critical to drawing value from efforts in task prediction?** If the goal of task prediction is to achieve something beyond confirming or denying Yarbus' assertions, there is value in ensuring that methodology allows for analysis beyond comparing prediction accuracies. We therefore employ methods for task prediction that are amenable to considering the relative importance

of associated gaze and image related statistics, and discuss additional considerations of importance at the level of methodological details.

3. **Which gaze statistics are most important?** In considering a set of relatively coarse grained task directives and choosing suitable methods, we aim to establish which gaze statistics or image derived features, including combinations of features, seem to diverge most across different task definitions. This provides insight into information represented within different types of gaze or image related statistics.

4. **What does task prediction tell us (and not tell us) about vision?** Task prediction establishes that different tasks may be distinguished on the basis of gaze statistics. Results presented in this thesis further reveal the relative importance of different types of features for the coarse-grained tasks examined in this thesis, and also within existing studies. However, it is important to address the limitations on what studies in task prediction are able to convey about human vision. An additional goal of this thesis is therefore to establish what benefits and limitations exist in examining task specific behaviour within a task prediction paradigm.

The balance of the thesis is structured as follows: In section 1.1 we present a survey of studies that emphasize task prediction highlighting differences in the set of tasks considered, methods and accuracies achieved across these studies. Chapter 2 presents the analytical methods that are exercised in this thesis. This includes further details on the datasets, and types of features considered for task categorization. Following this, in chapter 3, we first consider a dataset defined by explicit tasks assigned to observers and associated gaze data in regards to the relative classification performance in predicting task. This is achieved in considering the spatial distribution of fixations, local features at fixated locations, fixation

dynamics and global scene structure. This analysis considers different conditions, including pooled fixation data across all observers, as well as fixations for single observers subject to different methods of partitioning the image set. Various combinations of 4-way, 3-way and binary classifications are considered where appropriate to shed further light on factors that separate tasks. The second dataset's results in inferring user affective responses are contained within chapter 4. We discuss the broader implications of this analysis in sections 3.2 and 4.3 respectively, including limitations and possible fruitful directions forward. Finally, chapter 5 summarizes important results from this thesis in addressing the role of task in observed fixation behaviour, and presents fruitful directions for future work.

1.1 Prior work in task prediction

There are a variety of recent studies that consider how gaze interacts with task with direct reference to Yarbus' work, or specifically involving a classification paradigm for assessing the predictability of task from fixation data. DeAngelus and Pelz [2009] re-examined Yarbus' work, including tools, methods, and implications of Yarbus' findings. They also replicated Yarbus' original experiment using updated methods and a larger pool of participants and paintings for fixation recording. Their results demonstrated patterns consistent with Yarbus' data for Repin's painting using modern eye tracking devices, while restricting observations to a shorter time course. Castelhano et al. [2009] showed that an observer's task (object search and memorization) influences eye movement behaviour at the level of fixation durations and saccade amplitudes, specifically at the level of aggregate eye movement measures rather than individual fixation or saccade statistics. In the case of the memorization task, a larger area of the image was fixated and while the average fixation duration did not vary significantly

between tasks, certain areas were re-fixated (approximately) increasing their total fixation duration. These results were observed in viewing photographs of natural scenes for an object search task and a scene memorization task. Mills et al. [2011] also examined the impact of task on spatio-temporal fixation statistics with similar findings. A different task set was used in this case with free-viewing and *pleasantness* tasks added. The memorization task focused on the scene rather than objects, and the search task involved finding an "N" or "Z" that had been added to the scene, rather than a contextually appropriate object. While many of the results of the Mills et al. [2011] study mirror those of Castelhano et al. [2009], the average fixation duration varied more between tasks, specifically in the first few seconds of observation. The authors attributed this discrepancy to subtle differences in task design. Tatler et al. [2010] provide a detailed examination of Yarbus' body of work and the historical context surrounding it. Following this biographical overview an experiment was performed using a photograph of Yarbus. Results demonstrated that task influences the features fixated in viewing faces, and also that the importance of task in viewing extends to specific types of visual stimuli.

Greene et al. [2012] present a study suggesting that the relation between task, and fixation behaviour may be overstated based on an inability to predict task from the fixation data captured for Yarbus style experiments. The feature vector used by Greene et al. [2012] was composed of summary statistics of the observers fixations: (1) number of fixations, (2) the mean fixation duration, (3) mean saccade amplitude, and (4) percent of image covered by fixations assuming a one-degree fovea. These statistics had been used in previous work on scanpath analysis [Mills et al., 2011; Castelhano et al., 2009]. Also considered, was the proportion of fixation duration on various regions of interest: (5) faces, (6) human bodies, and (7) objects. The results from the Greene et al. [2012] study demonstrated

the ability to identify both the image and observer identity but an inability to predict the corresponding task. Subsequent efforts using this same data have demonstrated that while this presents a challenging classification task, above chance scores are possible with careful feature selection [Borji and Itti, 2014], division of data [Kanan et al., 2014], or using Probability Density Functions and Hidden Markov Models to model dynamics of scan-paths and latent contributions of covert attention [Haji-Abolhassani and Clark, 2014].

Borji and Itti [2014] re-considered the Greene et al. [2012] data, but chose to include low-resolution fixation density patterns as part of the evaluated feature set. This resulted in above chance performance of 34.14% accuracy across 4 tasks (25% chance-level). They also conducted a second experiment, mirroring Yarbus' original 7 task experiment, with prediction results of 24.21% when considering 7 tasks (14.29% chance-level). Their improvement in task classification accuracy by nearly 10% above chance was achieved principally by revising the feature set and inference methods used. Kanan et al. [2014] approached the Greene et al. [2012] data by first reproducing the original experiment with similar results. To preserve the temporal information for eye movements a Fisher Kernel Learning (FKL) algorithm was used, that allowed the variable number of time-series statistics to be compressed into a single feature vector. They also performed task prediction using two within-subject prediction experiments. For the first, this involved leave-one-out cross-validation using 19 of the 20 trials to train the SVM, and testing with the remaining single trial, was repeated for each possible leave-one-out combination (20 total). In the second condition, they trained the SVM using 4 of the 5 trials for each task, and used the remaining 4 trials (1 per task) as a test set, repeating this for each combination (625 total). Due to the increased number of training and testing trials to consider in the second approach, the number of states in the FKL was decreased from 10 to 5. The first approach resulted in a prediction accuracy of

52.9% (25% chance-level) using the FKL algorithm, while the second approach returned a lower 34.1% (25% chance-level) accuracy.

In more recent work, Haji-Abolhassani and Clark [2014] focused their attention on the Greene et al. [2012] data, having previously demonstrated good prediction accuracies on synthetic images [Haji-Abolhassani and Clark, 2011a,b]. An important contribution of their work is the use of Hidden Markov Models to estimate hidden state information from observed fixations to simulate the role of covert attention. The authors noted that in many cases the Centre of Gaze (COG) does not match the Focus of Attention (FOA). This was motivated by the observation that fixations did not always land on task specific objects, but observers where shown to be aware of the "overlooked" objects. To account for covert attention, the authors used a Gaussian Mixture Model to capture task relevant spatial positions and define the observation likelihoods via probability density functions. Probability densities correspond to different states derived from task-relevant objects/regions. To apply this to the Greene et al. [2012] data, they used k-means clustering on the aggregated fixations for each image-task pair, to generate a set of regions likely to be task-relevant. Using this approach, they were able to achieve a prediction accuracy of 59.64% (chance: 25%)

Coco and Keller [2014] also revisited the Greene et al. [2012] data, hypothesizing that the tasks considered may have required only visual processing rather than including other cognitive modalities such as language processing. By their hypothesis, similar processing requirements should result in similar strategies for allocation of attention, leading to harder inferences. To verify that tasks requiring distinct cognitive processes can be classified with greater ease, Coco and Keller [2014] carried out an alternative experiment. Tasks chosen for the experiment were: Visual Search, Object Naming and Scene Description. Each of these tasks required a different mixture of visual and language processing. Each task was

shown to a different group of participants, 25 for search and scene description each and 24 for object naming. The authors used the same 7 features as Greene et al. [2012] as well as another set of 15 corresponding to temporal fixation measures. Along with in-depth analysis of the recorded eye movement features across the various tasks, the authors also trained 3 different types of regression models to predict the observers tasks: multinomial regression, least-square angle regression, and support vector machines. Their accuracy averaged over all 3 tasks was 81% (33.33% chance-level). They also produced classification accuracy of 76% using only the features from Greene et al. [2012].

In contrast, the data presented by Henderson et al. [2013], includes 196 natural images and 140 images of text, and 4 tasks consisting of search, memorization, reading and *pseudo-reading*. Task performance in this case was approximately 80%, likely owing in part to the distinct nature of the chosen tasks. In addition, the mix of natural scenes and text involves very different stimuli, implying the possibility of stimulus driven differences as opposed to principally task driven differences [O'Connell and Walther, 2012].

Bulling et al. [2013] have presented a system to infer high-level contextual cues called EyeContext. Four participants were fitted with mobile eye tracking equipment consisting of 5 electrodes centred around the right eye, and were tasked with self-annotating various cues encountered during the day. The 4 categories participants were asked to track were: *"social (interacting with somebody vs. no interaction), cognitive (concentrated work vs. leisure), physical (physically active vs. not active), and spatial (inside vs. outside a building)"*. The eye movements recorded by the electrodes were encoded into fixed-length words composed of symbols, e.g. with a saccade to the left represented by the character 'L' and diagonal right represented by a 'B'. A sum total of 42.5 hours of data was collected across the 4 participants. They trained a string kernel SVM with 70% of the data and tested classification using the

remaining 30%, using a 5-fold cross-validation. The mean precision and recall obtained were 76.8% and 85.5% respectively. This research illustrates a novel way that task inference and eye tracking can be approached as this set of experiments more closely resembles real-life tasks and conditions.

A related effort considers the problem of recognizing the type of document an observer is currently reading. Kunze et al. [2013] performed an experiment involving 8 participants situated in 5 different environments and reading 5 different document types. This is implicitly a task inference prediction, as the different documents types are likely to elicit behaviours typical of highly learned document dependent strategies for reading. To track eye movements in various environments, observers were fitted with eye tracking glasses. In each case, 10 minutes of document reading was recorded and subsequently divided into 1 minute windows, as input to a decision tree (C4.5/J48). On a per-observer basis the decision tree achieved a accuracy of 99%, whereas average accuracy was 74% when observer identity was unknown and all observer data was pooled. Prediction accuracy increased to 90% for this latter condition when using majority voting over the entire 10 minutes of captured data. As noted by the authors this research could be used in reading assistance and logging of reading activities, and may also be expanded to other non-reading tasks.

Cerf et al. [2009] investigated the power of saliency maps to predict which image an observer fixated using only their scanpath data. While they achieved good results by combining the scanpaths from all observers, superior results were achieved when individual observers were factored into the prediction. They also proposed a metric to quantify the "Decodability" of datasets from certain feature sets (e.g. scanpaths), which may allow for the clustering of observers based on the features appealing to each individual. One of the uses of such features involves separating samples from special patient populations from those of control observers.

This was investigated by Tseng et al. [2013] who distinguished children with attention deficit hyperactivity disorder (ADHD) or fetal alcohol spectrum disorder (FASD) from a control group using a variety of features with an emphasis on saliency-based features. They also found that elderly patients with Parkinsons disease (PD) could be identified using primarily oculomotor related measurements. In related work, Jones and Klin [2013] studied infants from the age of 2 to 24 months and noted that infants that were later diagnosed with autism spectrum disorders (ASDs) exhibited a mean decline in fixations of caretakers eyes between 2 to 6 months of age.

In this thesis, we employ Random Forests [Breiman, 2001] for classification, and the details associated with this process are further explained in section 2. The reason for choosing Random Forests for classification in our work, is the capacity to identify the value of different features in successfully predicting the observer's task, while also modeling interaction between different types of features. This presents the possibility to identify important, and sometimes subtle differences in behaviour corresponding to different tasks, from a vantage point defined by the features that are considered. Random Forests are also employed by Sugano et al. [2014] to predict which of two images presented side-by-side on a computer monitor is preferred by an observer, based on eye movements. Their experiment was split into two phases: In the first phase, 11 individual observers were shown 80 pairs of image with no instruction. This was followed by 400 pairs of images with instructions to explicitly choose a preference between the two via a manual key press. Finally, the original 80 pairs were again shown but with the preference instructions. A total of 25 features were computed from fixations and saccades and used to train Random Forests. The mean accuracy was 73% (50% chance) when tested on the 80 image pairs with explicit preference instructions. When testing on the 80 image pairs with no instructions (Free Viewing) accuracy was only 61%

(50% chance).

Important characteristics of efforts involving task prediction from gaze are summarized in Table 1.1.

Table 1.1: Past Task-Prediction Contributions

	Obs.	Size	Tasks	Features	Methods	Performance
Henderson et al., 2013	12	◇ 196 images ◇ 140 texts	◇ Search ◇ Memory ◇ Reading ◇ Pseudo-Reading	◇ Eye Movement Measures	◇ Multivariate Pattern Analysis	68%–80% chance: 25%
Haji-Abolhassani and Clark, 2013 (Ex1)	6	180 images	◇ Counting 6 object types	◇ Gaussian mixture model capturing task relevant spatial positions	◇ Hidden Markov Models	71.5%–88.5% Indv task predictions
Haji-Abolhassani and Clark, 2013 (Ex2)	6	26 images	◇ Spelling 3 letter words	◇ Gaussian mixture model capturing task relevant spatial positions	◇ Hidden Markov Models	75.8%–87.7%
Lethaus et al., 2013	10	70km Simulated driving	◇ Overtaking ◇ Following ◇ Overtaking mult.	◇ Fixation time distribution across 4/5 zones during 5/10 second window.	◇ Artificial Neural Net ◇ Bayesian Net ◇ Naive Bayes Classifier	Real-time: 85% 1s delay: 90%
Bulling et al., 2013	4	Recording of workday (∼10 hours)	◇ Social ◇ Cognitive ◇ Physical ◇ Spatial	◇ Eye Movement encodings	◇ SVM	Precision: 76.8% Recall: 85.5% chance: 25%
Kunze et al., 2013	8	5 books	◇ 5 document types	◇ Saccade direction counts, mean & variance ◇ 95% quartile distance ◇ Slope over fixations	◇ Decision Tree	Independent: 74% Dependent: 99% chance: 20%
Greene et al., 2012	16	64 images	◇ Memory ◇ Decade ◇ People ◇ Wealth	◇ Eye Movement Statistics	◇ Linear Discriminant ◇ Correlation Methods ◇ SVM	25.9% chance: 25%
Kanan et al., 2014	16	64 images	◇ Memory ◇ Decade ◇ People ◇ Wealth	◇ Eye Movement Statistics ◇ Cartesian Coordinates	◇ SVM	Within Subject: 37.9% chance: 25%
Haji-Abolhassani and Clark, 2014	16	64 images	◇ Memory ◇ Decade ◇ People ◇ Wealth	◇ GMM for task relevant positions determined by fixation clusters (k-means)	◇ Hidden Markov Models	59.64% chance: 25%
Borji & Itti, 2014 (Ex1)	16	64 images	◇ Memory ◇ Decade ◇ People ◇ Wealth	◇ Eye Movement Statistics ◇ Spatial Density	◇ KNN ◇ RUSBoost	34.14% chance: 25%
Borji & Itti, 2014 (Ex2)	21	15 images	◇ Yarbus' Original 7 tasks	◇ Eye Movement Statistics ◇ Spatial Density	◇ K-NN ◇ RUSBoost	24.21% chance: 14.29%
Coco & Keller, 2014 (Ex2)	24	24 images	◇ Visual Search ◇ Object-Naming ◇ Scene Description	◇ Eye Movement Statistics	◇ Multinomial Regression ◇ Least-square angle Regression ◇ SVM	81% chance: 33.33%
Sugano et al., 2014	11	480 img. pairs	◇ Free View ◇ Preference	◇ Fixation & Saccade Measure Statistics	◇ Random Forests	61% & 73% chance: 50%
Boisvert & Bruce, Under Review	19	800 images	◇ Free View ◇ Object Search ◇ Saliency ◇ Explicit Sal.	◇ Spatial Density	◇ Random Forests	69.63% chance: 25%
Boisvert & Bruce, Under Review	19	800 images	◇ Free View ◇ Object Search ◇ Saliency	◇ Spatial Density ◇ HOGs ◇ LM Filters ◇ Gist	◇ Random Forests	56.3% & 76.9% chance: 33% & 50%

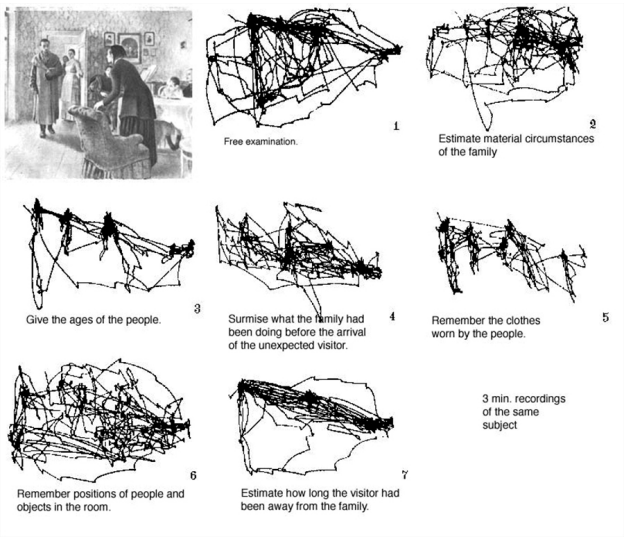

Figure 1.1: Examples of scan-paths, and questions posed from Yarbus' classic experiments. Examples of this form motivate the claim that task is influential in determining fixation behaviour. [Yarbus, 1967]

Chapter 2

Analytical Methods

In the following, we outline the details of the data considered in our analysis, classification methods and feature types considered for task prediction. In short, we first consider data tied to relatively general task directives including free viewing, object driven viewing, or saliency driven viewing. A second dataset consisting of affective response to magazine advertisements is also considered. Features for analysis include the spatial distribution of fixations, local oriented edge structure/contrast at fixated regions, holistic scene structure, and fixation specific statistics (e.g. saccade amplitudes). While a number of different classifiers have been examined, we have primarily considered random forest based bootstrap aggregation [Breiman, 2001] in our analysis. The reasons for this, details of features considered, and the specifics of the dataset are further discussed in the remainder of this section.

2.1 Fixation Data

2.1.1 Predicting Task

We first examine the data from Koehler et al. [2014], which includes fixation data for a set of 800 images associated with a number of different tasks. For each of the 800 images, each participant was asked to perform 1 of 3 viewing tasks (Free Viewing; Object Search; Saliency Viewing) or an explicit judgement task that required identifying the most salient location in the image through manual selection. A minimum of 19 (19-22) participants performed each viewing task and 100 participants performed the explicit judgement task given that each sample from participants yields only one observation per image for this task. In the Free Viewing Task, observers were instructed to freely view the image with no further direction given. For the Object Search task, the participants were given the name of an object to find in the scene, with the object present in 50% of cases shown. In the Saliency Search task, observers were asked to judge whether the left or right half of a scene contained the most salient region. Finally, in the Explicit Judgement task, observers were asked to use the computer mouse to click on what they believed was the most salient point in each image. Fixation data was collected using an EyeLink 1000 System at 250 Hz. For more detail on the precise methods, the reader is referred to the original experimental description [Koehler et al., 2014]. Fixation data for all three of the viewing tasks, and the Explicit Judgement data is used in our analysis. The explicit judgement data was only used in one of the prediction studies presented in this thesis, since the distinct nature of this data and absence of spatial fixation bias makes discriminating this task from the gaze based tasks relatively easy. In addition, given only one data-point per image for the Explicit Judgement case, the only natural comparison using this data is in aggregating data across all participants. Therefore,

for the aggregate classification case, the fixation data for each image/task combination is pooled across all observers. In contrast, the *individual classification* cases that make up most of the analysis in Chapter 3 consider each image-participant-task sample as a separate case. It is important to note that in tests of classification performance, our results consider explicitly the role of prior exposure to the test image set during training, providing additional observations relevant to task prediction in a broad sense.

2.1.2 Predicting Affective Response

The second dataset considered in this thesis is based on Observer Rating data from Lundqvist [2015] which is comprised of 108 images of Swedish ads from which 8 different subjective ratings were generated by users, resulting in a total of 2735 Image-Observation pairs. The 8 ratings are: Brand Experience, Interest, Understanding, Relevance, Valence, Memory, Purchase Intention, and Originality. We were unable to perform analysis based on the observers as the information pertaining to participants was anonymised and data collection spanned multiple days, resulting in an asymmetric distribution of observations, but this should not affect the analysis in other categories. In addition to the prediction accuracy, we also examine the correlation between the different categories of ratings and attempt to identify independent axes in the raw rating data that retain the vast majority of the variance present in the raw data samples. The Joint Approximate Diagonalization of Eigen-matrices algorithm (JADE) algorithm by Cardoso [1994] and it's extension Cardoso and Souloumiac [1993] are used to this end. This is a blind Signal Source Separation algorithm, and is similar to applying factor analysis, without any assumptions required about the data. Affective responses that originate from common hidden/latent sources are pooled onto a common axis JADE, and affective dimensions are as independent as possible. This allows another means

of illustrating the relationships that exist between user data and gaze. The source signals identified by JADE can also be used to reduce the computational complexity required to predict the responses for a novel observer in lowering the dimensionality of the response vector with principal component analysis (PCA) a preprocessing step in JADE.

2.2 Features

A central goal of this thesis is evaluating both the role of spatial position (or density) of fixations in task prediction, but also the diagnostic value of specific features or image structure at fixated locations. We have also examined whether scene structure combined with the spatial distribution of fixation patterns might provide a useful diagnostic. To this end, we have selected a number of features for testing task classification performance. Each type of feature is considered in isolation and also in various combinations with other types of features.

It should be noted that our use of the term "features" is different in meaning from what is typical in the Psychology literature, wherein a feature tends to refer to the nature of stimuli and their characteristics (e.g. angles, shape, motion, etc.). In this thesis the term "features" refers to data or statistics that are extracted from an image or gaze recordings and is much more general. These features are used as input for the inference model.

2.2.1 Fixation Density Map (Density)

An alternative to considering raw fixation positions given their relative sparsity (in pixel terms), is considering a more continuous density map derived from the raw fixations. This may be produced by way of convolution of the fixation map with a Gaussian profile [Bruce and

Tsotsos, 2006], and/or sub-sampling to produce a coarse-grained continuous spatial density map for task prediction [Borji and Itti, 2014]. The first set of features we have considered are spatial densities of fixations. We represent this quantity by generating a density map of the fixations on the images. A continuous density map is produced by convolving the image with a 2D Gaussian envelope with standard deviation corresponding to 1 degree of visual angle (27 pixels). These density maps are then down-sampled by a factor of 15 for analysis involving task prediction, which results in a 27x27 map or a 1x729 feature vector. In predicting affective response values the density map was down-sampled by a factor of 32 which results in a 32x24 map or a 1x768 feature vector.

Predicting Task: Aggregate Density Map (Aggregate Density:) We also tested an aggregate density map approach, where the fixations from all observers (as opposed to single observers) for a single image were merged to form a density map. That is, for image number $i \in \{1, ..., 800\}$ and task $j \in \{1, ..., 3\}$ fixations across the 19 observers were aggregated into a single density map. For the explicit judgment task, the aggregated data is based on pooling of the Explicit Judgements across 100 participants. This yields a number of observations per image similar to the fixation data. This pooling of the fixation data serves primarily to facilitate a comparison between data from viewing tasks, and from the explicit judgment task.

2.2.2 The Leung-Malik (LM) Filter Bank

The LM filter bank [Leung and Malik, 2001] consists of 4 Gaussians filters corresponding to different spatial scales, 8 Laplacian of Gaussian (LoG) filters at different spatial scales, and first and second derivatives of Gaussians for each combination of 3 spatial scales and 6 different orientations (36 additional filters) for a total of 48 filters. To create an LM based

feature vector, each image was convolved with the LM-Filter bank at it's original scale, and the response of the 48 filters was sampled at each fixated location. Given that the total number of fixations is variable for any given image (generally from 7-15 fixations), filter responses across all fixations for an image were converted to summary statistics given by the mean response, and the standard-deviation of the response for each filter type across all observed fixations. This produces a 96 dimensional feature vector that captures the mean response, and variability in response of each of the filter channels for fixated regions of an image. It is worth noting that the LM filter bank is chosen in part for its similarity to model simple cells represented within V1 and characterized by Gabor-like and center-surround receptive field profiles. This choice is relevant to making stronger assertions about human vision through task analysis by classification, and is an important consideration for future efforts in task prediction going forward.

2.2.3 Histogram of Oriented Gradients

Histograms of Oriented Gradients are widely used in the computer vision literature [Dalal and Triggs, 2005], and have shown success in a range of tasks including object detection [Felzenszwalb et al., 2010] or scene classification [Xiao et al., 2010]. The HOG descriptor consists of histograms corresponding to oriented edge structure at different spatial scales within a local window of the image. Such features therefore capture coarse-grained summary statistics on the distribution of angular and radial frequencies represented within a local region of an image. While it's evident how such a representation may be used to determine whether an object is present at a given location, it's also natural to consider whether there is some inherent bias in edge content expressed in viewing behaviour across different tasks. Fixation based HOG features were generated in a fashion similar to the LM filter bank: At

each fixated location in the original image, HOG features are extracted corresponding to a 65x65 image patch centred at the fixated location [Felzenszwalb et al., 2008]. This results in a 31-dimensional feature vector for each fixation. Again, given variable numbers of fixations, the 31 dimensional HOG feature vector was converted to a summary representation, in considering the mean and standard deviation of HOG features across all fixations, yielding a 62 dimensional feature vector.

2.2.4 Number of Fixations

One of the most common features used in task prediction, the number of fixations has been shown to vary based on tasks in a number of previous papers [Castelhano et al., 2009; Mills et al., 2011; Tatler et al., 2010]. We have also found the number of fixations to be particularly useful in combination with the LM Filters and HOGs which capture summary statistics of image content at the location of fixations.

2.2.5 Scene Gist

We have also considered a representation that captures the holistic structure of a scene based on the Gist descriptor [Oliva and Torralba, 2006]. The Gist descriptor is produced in sampling the responses of local filters sensitive to intensity gradients at different spatial scales, and over a grid of sub-windows on the image. These are subsequently converted to low dimensionality holistic receptive fields through PCA. This representation has been demonstrated as capable of classifying the type of scene (indoor, outdoor, forest, city, etc.) [Oliva and Torralba, 2006], and also having use in improving performance of models for predicting gaze locations [Torralba et al., 2006]. The motivation for this set of features, is to examine whether general holistic scene structure is able to augment the ability to predict

task when coupled with spatial densities of fixations. This is motivated by the notion that the spatial envelope of a scene might play an important role in defining fixation patterns, independent of task.

2.2.6 Feature Combinations

The relative value of individual base features types is important, but we are also interested in additional value that may be had in leveraging multiple distinct feature types for prediction. To this end, a number of composite feature sets have also been considered based on various combinations of spatial fixation density maps, fixated image features, scene structure and dynamics. One motivation for the incremental composition of features in our analysis, is in establishing a more detailed understanding feature importance. The incremental gains in evaluating different subsets of features, provides additional information on redundancy in information captured by different feature sets with respect to the statistics that define task boundaries.

2.3 Classifiers

For results presented in this thesis, Random Forests [Breiman, 2001] are used for classification[1]. In all cases when Predicting Task, 50% of the data was used for training, and the other 50% for testing. Given the smaller dataset available for predicting affective response, the training and testing sets were defined as 80% and 20% of all samples respectively. Classification is performed on different combinations of spatial and/or structural features to assess the relative efficacy of different cues and to determine prediction performance.

Random forest classification relies on the consensus predictions of a number of distinct

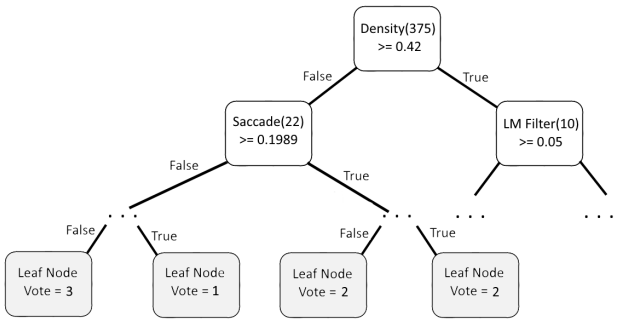

Figure 2.1: A sample of decisions that may be taken by the nodes within a decision tree, the resulting leaf node corresponds to the tree's vote. Votes are collated across all trees comprising the random forest to determine decision of the ensemble.

decision trees. Each decision tree comprises a hierarchical arrangement of decision nodes. For example, the learning process might result in a root node that passes on an observation to one child node if the spatial fixation density for a particular location in the image is above some threshold, and to the other child node if the density is below this threshold, as depicted in Fig. 2.1. After a series of decisions of this type, which may also include branches on nodes that include local image features (HoG, LM), contextual statistics (Gist), or fixation statistics (number of fixations, saccade amplitudes), a leaf node is eventually reached that indicates the predicted task. An effective strategy for classification is to use a number of such decision trees in concert with an overall classification decision based on a majority weighted vote from individual decision trees. This can help to control against overfitting the data, but also brings additional benefits in diagnosing the value of features. In generating a collection of decision trees that make up a random forest, each decision tree is produced from independent data samples. Data samples from the training set are selected with replacement to produce a unique training set for each individual tree. Samples from the training set that

are not included in a given sample are referred to as *out of bag* samples. Performance for a decision tree on *out of bag* samples provide one useful characteristic for understanding sample and feature importance. In the training process, when considering features to branch on in the tree, only a subset of the total variables/features is considered. This also introduces additional randomness that serves to produce diversity in the structure of trees that make up the random forest. In our evaluation, these bootstrap samples for decision nodes were based on $\sqrt{N}$ samples, where N is the total number of statistics (features) used by the classifier. Factors that have a greater impact on prediction performance have more diagnostic value and are more important features in separating different task categories. A determination of feature importance may be made by permuting the values of a particular feature across different data samples. For example, the mean response of one of the LM filters across fixations provides a predictive statistic for each image. If these values are shuffled across the samples, it is possible to measure the impact on performance. This is performed for out of bag samples, providing a measure of feature importance for all of the individual features. This is an important property of this classification strategy, as it brings the additional value of discerning relative feature importance to understanding task-feature relationships. The detailed mathematical justification for this analysis is beyond the scope of this work, but the interested reader may refer to the statistical motivation given by Breiman [2001].

[1] Alternative classification methods were also evaluated. The rationale for testing an array of classification methods was to confirm that results are representative of features chosen for prediction, and not the choice of classifier. All other classification methods performed no better than Random Forest based classification,and Random Forest based classification was more stable (similar performance using alternative methods sometimes required careful choice of parameters).

1. **Neural networks**: This evaluation employed multi-layer neural networks comprised of 2 or 3 layers of sparse autoencoders [Vincent et al., 2008], with the weights of the sparse autoencoder training used to initialize a standard 2 or 3 layer back-propagation [Vogl et al., 1988] network for classification. Performance was not any different than using Random Forests for the best cases, however sensitivity to parameters resulted in much poorer performance without careful tuning.
2. **Lasso regression** [Tibshirani, 1996]: Regularized L1 Logistic regression was also evaluated for classification performance. With an appropriate choice of λ, results were on par with accuracy using

In the Predicting Task chapter, a range of values was considered for the total number of voting tree-based classifiers including 50, 100, 200, 500, 1000, and 2000 trees respectively. This is in consideration of determining a ceiling on classification accuracy, but also provides confidence on the stability of the classification method used in much of the analysis. That is, a variable number of trees are considered to ensure that performance differences are due to feature differences rather than the complexity of the classifier.

Random Forests, however determination of this value was non-trivial, in part because this was dependent on features used for classification. In addition, mixed feature sets required normalization to achieve equivalent performance.

3. **AdaBoost** [Freund et al., 1996]: Several variants of AdaBoost were also tested. This method was relatively easy to obtain good prediction results, although in most cases accuracy values fell somewhat short of those produced by Random Forests. Among adaptive boosting methods, performance was best for the pseudo-loss variant of the standard algorithm [Freund et al., 1996] (as compared with LPBoost, TotalBoost and RUSBoost).

Chapter 3

Predicting Task

3.1 Results

In the following, we examine the performance for the Random Forest classifier across the various feature sets, and in considering various combinations of features. This has been examined for the aggregate case (all observer data pooled), as well as the individual cases. In addition, we also examine multi-way classification and pair-wise binary classification to examine the separability of different task directives based on the fixation data. The following demonstrates that most of the features considered present significantly above chance performance. This is notable due to the results of Greene et al. [2012] which questioned the basic assertion that task may be predicted from gaze. It is worth noting that accuracy is dependent on the division of images among training and test sets.

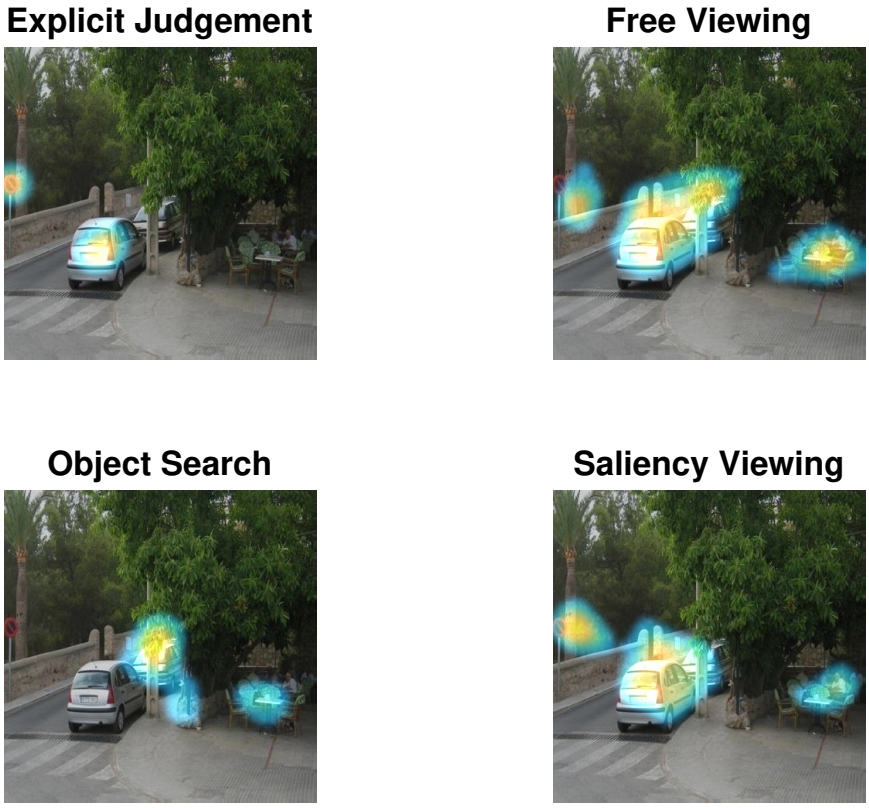

Figure 3.1: A visualization of fixation density for the aggregated explicit judgment data (100 participants) and observer fixations (19 participants) for a sample image. Note that similarity between Free Viewing and Saliency Viewing appears to be stronger than the Object Search and Explicit Judgement cases.

3.1.1 Aggregating Observers

Aggregating data across observers for each image/task combination results in 3200 (4*800) image / task pairs. Half of these instances were used to train a Random Forest while the other half were used for testing only. A depiction of aggregated fixations can be seen in Fig. 3.1, the distinction between tasks is evident though we do notice strong similarities

Figure 3.2: A sample subset of images representative of the Koehler et al. dataset. Koehler et al. [2014]

between Free Viewing and Saliency. A larger set of representative image samples from the same dataset may be seen in Fig. 3.2. Only the Spatial Densities were used for the aggregate classification test case as the resulting prediction accuracy is strong based on spatial bias alone, and this is the only type of analysis that is sensible for the Explicit Judgement task. Given that the Explicit Judgement task does not include any gaze data we cannot include some the other features in our analysis and aggregating observers is the only means of including the Explicit Judgement task in our analysis. More detailed analysis is there-

fore reserved for the 3-way classification case, where the explicit judgement data across 100 observers is not included, and the analysis is restricted to fixations. Given that the spatial density features have a topographical organization, it is possible to visualize the importance profile in a topographical layout (Fig. 3.3). It is evident that the central region is of greatest importance, and that the degree of spatial (central) bias, is one factor that has diagnostic value in determining the task being performed. This is consistent with many observations of the prominent degree of central bias present in gaze patterns, although this also hints that the extent and shape of central bias is task variant. Results for variable numbers of trees appear in table 3.1 revealing the relative stability as a function of number of trees. Subsequent results for individual observers present only results corresponding to 2000 trees given this relative stability.

Table 3.1: Aggregate Density Map Results (Prediction Accuracy)

Trees	All	Free/Obj	Free/Sal	Free/Exp	Obj/Sal	Obj/Exp	Sal/Exp
50	70.87%	84.38%	66.13%	88.25%	89.88%	97.50%	90.00%
100	69.31%	83.25%	65.75%	89.75%	89.75%	97.75%	89.75%
200	69.87%	83.63%	65.38%	89.38%	90.00%	97.62%	89.88%
500	68.87%	84.13%	65.75%	89.25%	89.75%	97.50%	89.88%
1000	69.44%	83.50%	66.75%	89.38%	90.00%	97.75%	90.12%
2000	69.19%	83.87%	66.37%	89.50%	90.25%	97.62%	89.62%

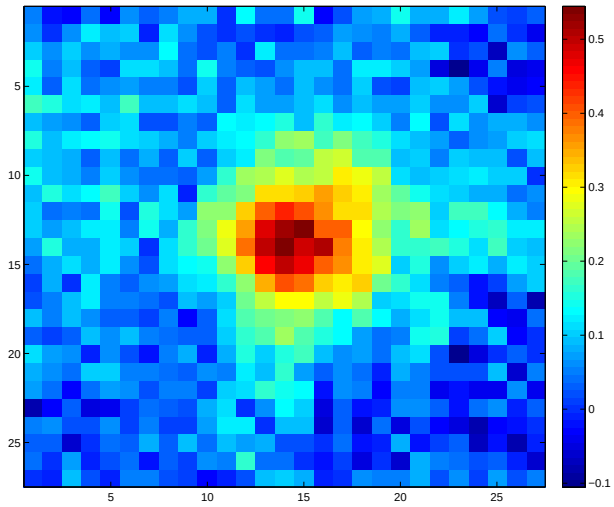

Figure 3.3: Importance of spatial density statistics to classification performance based on Aggregated explicit judgment coordinates, and fixations across all participants. Importance measures are arranged topographically corresponding to their position in the fixation density map. This indicates that the relative importance of observations follows a concentric spatial profile.

3.1.2 Individual Observers

For task prediction based on data from individual observers, we consider the first 19 observers across the 3 task conditions to equate the number of observers per task. In total, there are 19 observers x 3 tasks x 800 images for a total of 45,600 cases to be distributed among training or testing. To provide deeper insight into the task prediction problem in general, we have considered two different partitions of this data as follows:

Partition I: All of the observers, and all tasks are represented for half of the images. This implies that classifier predictions are not based on any patterns specific to individual images seen in training.

Partition II: All observers and images are represented, but only half of the tasks carried out by each observer appear in the training set, and the other half in the test set (with equal number of samples of each task in training and test sets). This allows the relative importance of the specific images used in training, and importance of the size of the image set to be discerned. These two data partitions are referred to as P(I) and P(II) from hereon.

Classification results corresponding to the various feature / task combinations are summarized in Table 3.2 for P(I), and in Table 3.3 for P(II). There are some notable difference in the efficacy of different features, and also a significant impact on accuracies as a function of how data is partitioned. These points are discussed in detail in what follows, along with careful analysis of diagnostic measures of feature importance.

3.1.3 Spatial Density

Similar to the case of classification based on aggregated observers, we assess the relative importance of different locations in the spatial density map (Fig. 3.4). The individual Spatial Densities shown in Fig. 3.4 also demonstrate significant weight on the importance of centrally located positions in the density map, albeit the individual case is characterized by a more pronounced peak at the very centre, accompanied by a more diffuse spread of feature importance over the scene outside of the centre. This again points to the importance of eccentricity of fixations as an importance distinction between tasks. Given that by aggregating multiple observers together we also increase the number of central observations these take on more importance as they reveal the minute differences in distribution between tasks. It is also interesting to note that relative importance reflects both differences in spatial distribution and relative frequency reflected in density. The importance of spatial (center) bias within studies of fixation behaviour has been explored in detail, and the data presented by Tatler

Table 3.2: Individual Density Map Results P(I) - All observers, all tasks, 50% of images

Spatial Density	LM Filters	HoG Features	Gist	Sacc. Ampl.	Num. Fixations	All Tasks	Free. vs Obj.	Free. vs Sal.	Obj. vs Sal.
Chance						33.33%	50.00%	50.00%	50.00%
✓						48.34%	66.55%	58.89%	66.13%
	✓					42.98%	65.05%	52.32%	62.18%
		✓				41.06%	61.98%	51.36%	60.20%
				✓		54.54%	76.89%	56.69%	75.65%
					✓	50.65%	73.11%	54.81%	71.38%
	✓				✓	49.45%	73.23%	51.84%	71.26%
		✓			✓	49.73%	73.35%	53.52%	71.33%
✓			✓			48.47%	66.54%	58.80%	66.12%
✓	✓				✓	50.09%	69.94%	59.27%	68.09%
✓		✓			✓	50.26%	69.38%	59.17%	68.19%
✓	✓	✓	✓			51.59%	71.34%	59.53%	69.32%
✓	✓		✓		✓	50.34%	69.35%	59.00%	68.15%
✓	✓	✓	✓		✓	51.62%	71.00%	59.48%	69.12%
✓	✓	✓	✓	✓	✓	53.42%	73.41	59.59%	71.01%

Table 3.3: Individual Density Map Results P(II) - All images and observers represented, 50% of tasks per observer

Spatial Density	LM Filters	HoG Features	Gist	Sacc. Ampl.	Num. Fixations	All Tasks	Free. vs Obj.	Free. vs Sal.	Obj. vs Sal.
Chance						33.33%	50.00%	50.00%	50.00%
✓						54.6%	76.24%	58.22%	75.18%
	✓					42.89%	65.07%	51.96%	62.83%
		✓				43.79%	64.60%	52.19%	62.95%
				✓		54.45%	76.86%	56.69%	75.37%
					✓	51.04%	73.42%	54.65%	72.13%
	✓				✓	49.79%	73.56%	53.25%	71.58%
		✓			✓	50.39%	74.19%	53.70%	71.89%
✓			✓			54.74%	76.27%	58.17%	75.27%
✓	✓				✓	56.05%	77.82%	59.58%	76.28%
✓		✓			✓	56.16%	77.70%	59.58%	76.35%
✓	✓	✓			✓	56.91%	78.70%	59.81%	76.88%
✓	✓		✓		✓	56.11%	77.86%	59.59%	76.33%
✓	✓	✓	✓		✓	56.74%	78.50%	60.20%	76.82%
✓	✓	✓	✓	✓	✓	58.14%	79.98%	60.16%	77.85%

[2007] reveals a similar trend to these observations for two very similar task definitions but corresponding to a different set of image data.

Differences in prediction accuracy between P(I) and P(II) are striking. While spatial density alone is among the more effective features for task prediction when all images appear in training (P(II)), it's value is diminished significantly when training and test image sets are disjoint. This has important implications for how studies on task prediction are interpreted, especially in light of the relatively large image set associated with the data under consideration relative to prior efforts in task prediction. A more detailed discussion of the implications of this observation appear in section 3.2.

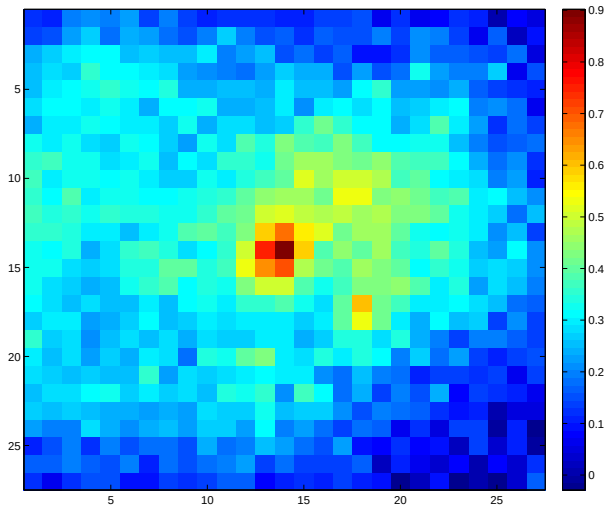

Figure 3.4: Importance of spatial density statistics to classification performance for individual fixation density maps (no cross-observer data aggregation). Importance measures are arranged topographically corresponding to their position in the fixation density map.

3.1.4 Local Image Features

The LM Filters and HoG features alone present similar efficacy for both P(I) and P(II). This corresponds to approximately 42% accuracy for the 3-way classification tasks, and approximately 52%-65% accuracy depending on task pairing, with free viewing and saliency viewing again most difficult to distinguish. In agreement with prior observations concerning the relative lack of importance of features at fixated locations, fixated image features alone are relatively poor at distinguishing between tasks. However, accuracy as high as 65% for some of the binary classifications suggests that statistical differences between these cases are not entirely spurious. It is possible that targeted psychophysics experimentation may yield further insight in this particular case.

Important to understanding this observation, is the relative importance of different fixated features to distinguishing tasks. To support this analysis we present the relative importance of different features from the LM filter set in Figure 3.5 based on out-of-bag analysis corresponding to the random forest based prediction. For illustrative purposes, first derivative filters corresponding to horizonal (red) and vertical (yellow) edge content are highlighted. There appears to be a consistent advantage to statistics associated with vertically oriented image structure at fixations in delineating task. A more detailed illustration of this difference is shown in Figure 3.6, which demonstrates the probability density associated with horizontal (left) and vertical (right) first derivative LM features for the 3 tasks. These correspond to features 7 and 10 appearing in Figure 3.5. Free viewing, object search and saliency viewing correspond to the blue, red and yellow curves respectively. The bottom row depicts the *boxed* part of these curves at a higher level of zoom to better show the separation among feature distributions. It is worth noting that these differences exist for the central part of the curve as well, but are more subtle. For vertically oriented structure, overall separation

is greater and also some degree of separation between the free viewing and object search conditions emerges. While this analysis does not reveal much about the reason for these differences, it demonstrates a subtle difference in structure of content at fixations that may be an important target for future analysis as it relates to task. It is also important to note that the value of this information is due chiefly to an accumulation of very weak evidence across a number of fixations in yielding task discrimination that is well above chance, and not any single fixation.

An interesting asymmetry appears in the relative importance of spatial density, and local image structure for binary task classification. There is an advantage for spatial density over local features in delineating tasks independent of the data partitioning scheme, but this difference is much larger for P(II). However, the value of spatial density is invariant to the data partition scheme for discriminating saliency viewing from free viewing. An implication of this, is that the spatial density profile for saliency viewing and free viewing differ in a manner that is relatively independent of content specific to individual images, while object search appears to carry a spatial density profile that is more highly image specific. This is also revealing with respect to the level of specificity of image patterns driving statistical differences across different tasks.

3.1.5 Global Image Features

While fixation density for most locations (especially those proximal to the center) is relevant to task discrimination, feature importance for Gist features approaches 0 for all Gist feature dimensions. One possible explanation for this is that the holistic spatial envelope has a relatively small influence on gaze targets insofar as it interacts with task. That is, the influence of holistic scene structure may be relatively strong overall, but task invariant.

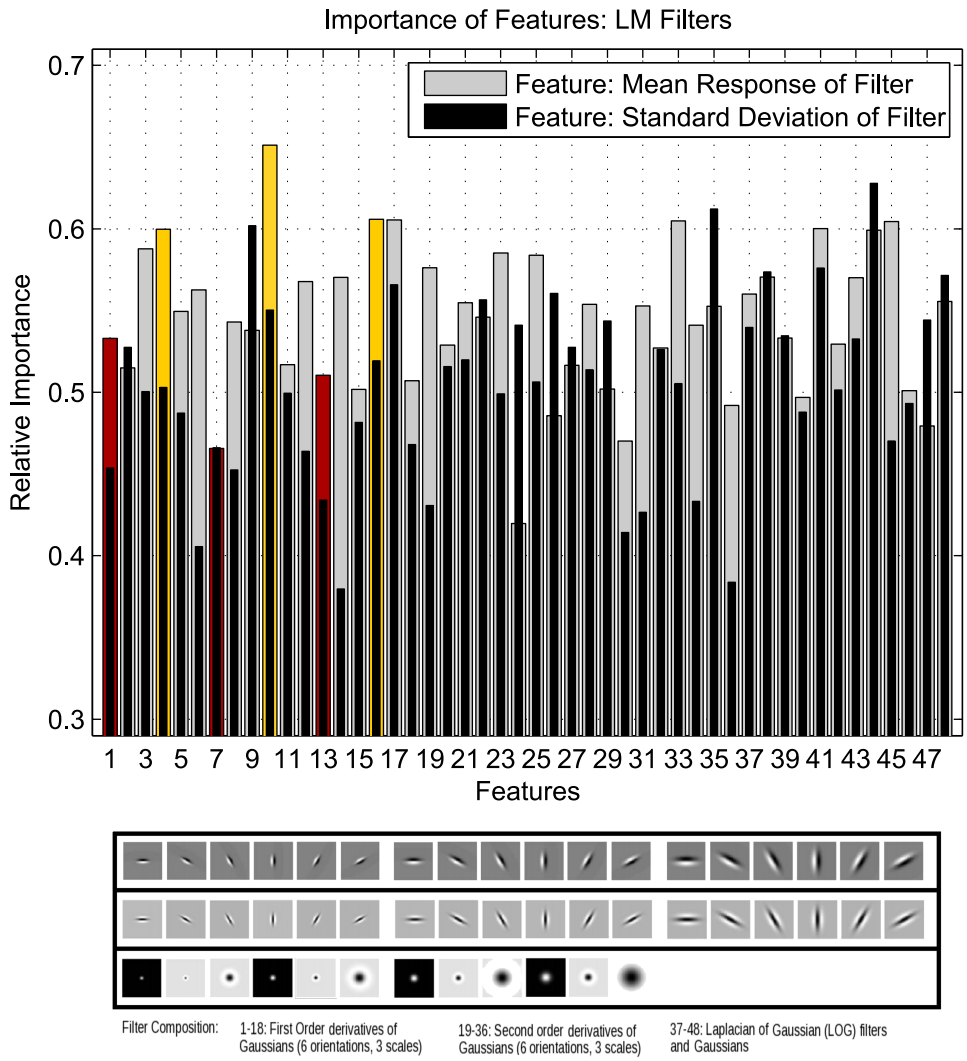

Figure 3.5: Feature importance for statistics corresponding to mean and standard deviation of LM filter outputs across fixations. The order of bars in the plot mirrors the order of filters in the legend below the bar plot in the following order: First row, left to right. Second row, left to right. Third row, left to right. First derivative filters show systematic variation as a function of orientation, with vertically oriented filters (yellow) carrying consistently higher predictive value than horizontal filters (red).

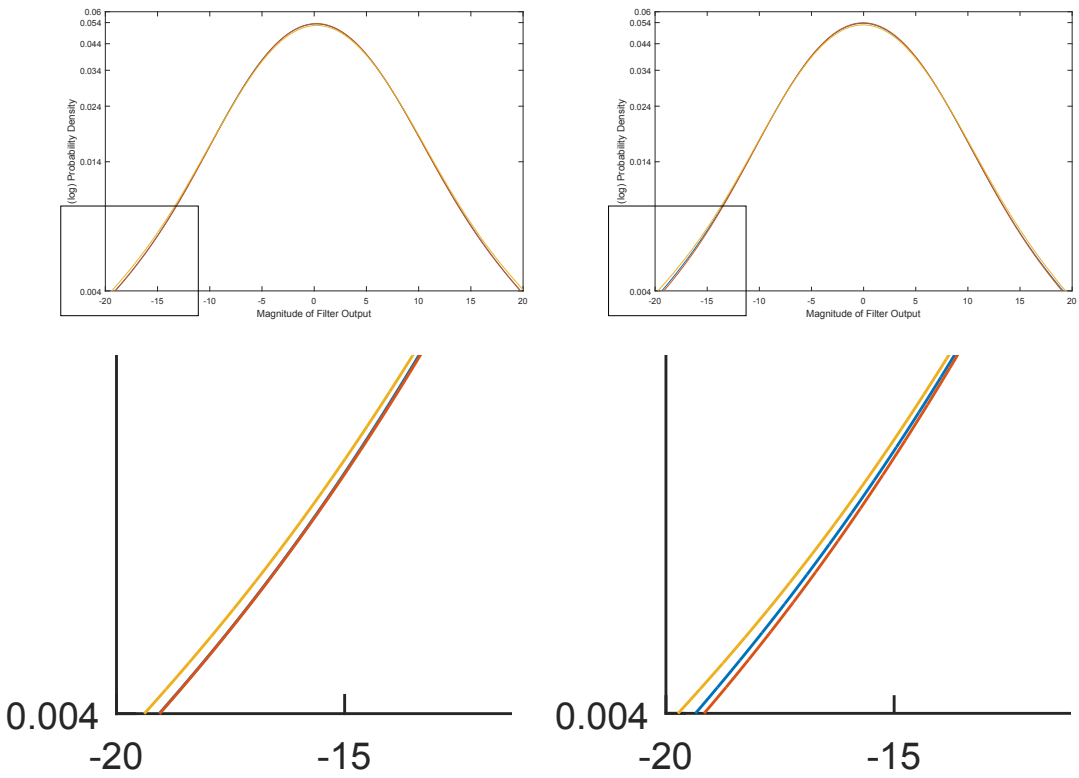

Figure 3.6: Probability density associated with the response of two different LM filters (see text for details). Colours correspond to response densities for Free Viewing (blue), Object Search (red), and Saliency Viewing (yellow). Density profiles are shown for the horizontally oriented first derivative of Gaussian LM filter (left), and for the vertically oriented first derivative of Gaussian LM filter (right). The lower frames depict a magnified view of the lower left section of each curve. Note that differences in separation of classes are present between the two plots across the entire curve, but these become more visually pronounced at the extremities.

An alternative possibility is that the task dependent influence of holistic/structural differences are already reflected implictly in the spatial density profiles. That is, the Gist based

structural representation may be a weaker cue when coupled with spatial density given redundancy in the information they capture. It is also the case that the number of image samples is small relative to what is typical for scene classification efforts [Torralba et al., 2006].

3.1.6 Fixation Dynamics

Fixation dynamics associated with the different tasks are characterized by the total number of fixations, and the amplitude of saccades observed within each task. These measurements are surprisingly effective in distinguishing between the tasks considered. Given that there are significant differences in total number of fixations (and latency) for the object search task compared with free viewing, and saliency viewing it is evident that this is a valuable statistic in distinguishing among these tasks. The probability density associated with the number of fixations for each class are shown in Figure 3.7.

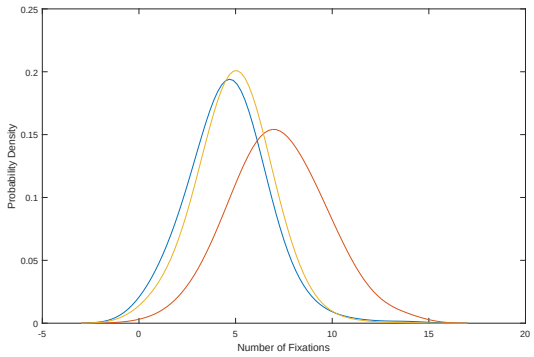

Figure 3.7: Probability densities reflecting the number of fixations made for each image presentation for Free Viewing (blue), Object Search (red) and Saliency Viewing (yellow) tasks.

Perhaps more surprising is the strength of saccade amplitudes alone in distinguishing

between tasks. In particular, for the challenging case of free viewing vs. saliency viewing, these are among the most important features alongside spatial fixation density. To examine this observation in more detail, we plot the relative importance of first, second and additional saccades made at the start of each trial in Figure 3.8.

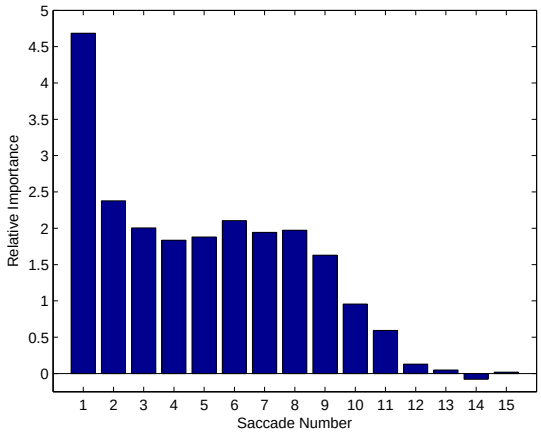

Figure 3.8: Relative importance values for the first several saccade amplitudes based on out-of-bag analysis corresponding to the Random Forest Classifier.

Figure 3.8 reveals that the amplitude of the very first saccade is highest in importance, but there is also a high value to the several subsequent saccades in discriminating between tasks. The probability density associated with saccade amplitudes is shown in Figure 3.9. This reveals that initially saccade amplitudes for object search are quite disparate from the other tasks, however, with an increasing number of saccades object search and saliency viewing converge, and distinguish themselves from free viewing. This observation is important in revealing the apparent value of fine grained temporal dynamics in providing defining traits associated with different tasks. Given that there exist differences between features at fixation for long versus short saccades that are task dependent, interaction between relative spatial position of saccades and content at fixation is also likely to be relevant to inferring task

[Tatler et al., 2006] even for relatively general tasks definitions such as those examined in this thesis. This also has implications for the role of task prediction for applications in human centric applications that make use of eye movements, with the assumption that there may be a significant advantage to classification models that employ a rich characterization of temporal dynamics.

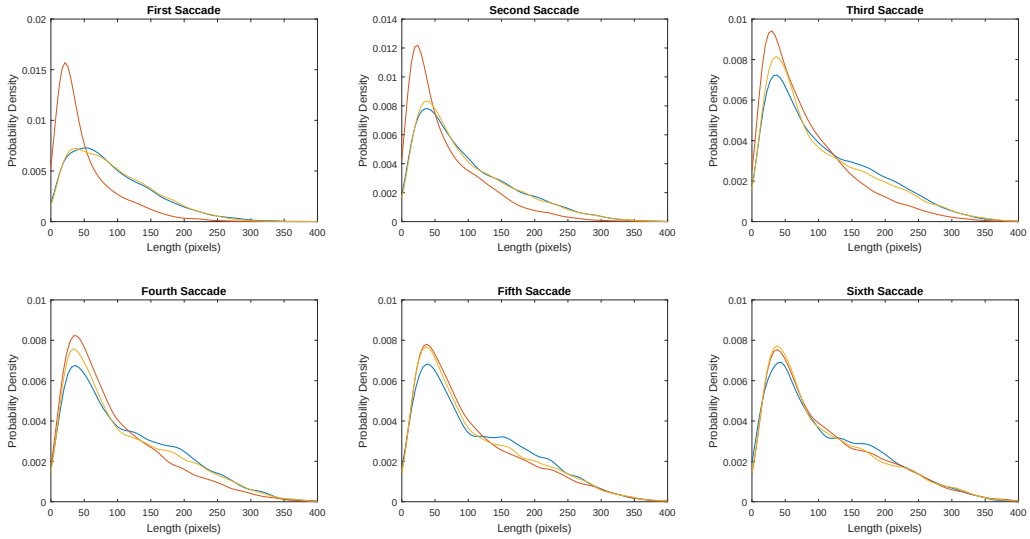

Figure 3.9: Probability densities associated with saccade amplitudes (in pixels) for the first 6 saccades. Colors indicate Free Viewing (blue), Object Search (red) and Saliency Viewing (yellow).

3.1.7 Feature Combinations

Performance for binary task classification is important to the overall interpretation of results as this provides a sense of similarity among behavioural observations among each pair of tasks. It is also evident in some of the preceding discussion within this section, that some important observations concerning relative importance of specific features or partitioning of

data are possible only with a granularity of task prediction results that includes different subsets of features. This includes determining the degree of redundancy in information captured by different types of features. For example, relatively small gains are observed in combining LM and HoG features compared to their independent prediction accuracies. Slightly larger gains result from combining spatial densities and fixated features, and even larger gains in combining saccade amplitudes with spatial density and fixated features. These observations largely fit with *a priori* intuition concerning the overlap in information represented among such features. However, such analysis also helps to support or rule out other suspicions concerning the nature of task differences. For example, one might posit that feature differences at fixation are due primarily to bias in the spatial density profile of fixations that varies with task in combination with bias in how images are composed (framed and targeted by the photographer). However, the improvement seen in combining these features seems to deny the possibility that feature level differences are entirely spatial in their impetus. It is important therefore to note the methodological value of decomposition of both features, and task pairing in prediction to derive a deeper understanding of task relatedness.

3.2 Discussion

We have considered the extent to which relatively general task definitions, such as free viewing versus search for objects may be distinguished on the basis of either the spatial density profile of fixations, features at fixation, scene structure or fixation dynamics. There are a variety of interesting observations that emerge from the classification experiments that include establishing the relative importance of different features in discriminating tasks, and highlighting important methodological considerations in how analysis by classification may

be conducted.

While free viewing and saliency viewing produce gaze derived statistics that are quite similar, these tasks are distinguished from object search with relative ease. Employing methods that place an emphasis on feature importance allows for subtle differences between tasks to be identified for more careful focused examination. For example, the relative importance of saccade amplitudes for the first several fixations in free viewing and saliency viewing may not be as readily observable using more traditional methods for statistical analysis of the data. In this view, some of the value of task prediction may reside in the capacity to test a large number of features in their value for task discrimination to identify targets for subsequent analysis using alternative methods. In this view, there is a role for task discrimination as a means for *high-throughput* screening for important feature dimensions for more careful scrutiny, or targeted psychophysics experimentation.

While a task description may carry a relatively clear intent or definition, the associated neural and behavioural mechanisms that any task definition elicits may be relatively obscured in comparison. With that said, there is reason to be optimistic that a further proliferation of studies focused on eye movements including a larger variety of task definitions will help to clarify this relationship. This will allow for a stronger functional characterization of the nature of different *tasks* at various levels of abstraction (high-level or specific), while also distilling out the distinct associated neural or behavioural mechanisms that are recruited for particular tasks. (For a different set of methods towards this goal, see also [Bruce, 2014]). Task prediction accuracies may be of value in determining task similarity, however, the specific features that are most discriminative in separating tasks may also provide important clues concerning how tasks are related. Finally, it is important to note that there are certain limitations to this type analysis and these are discussed in greater detail in section 3.2.2.

On the balance of evidence from studies that aim to predict task from eye movements, there is evidently support for Yarbus' assertions concerning the importance of task in determining gaze behaviour, and recent efforts have demonstrated that fixation data may successfully predict task. However, analysis that emphasizes feature importance is necessary to understand specific factors that distinguish tasks, and also to understand which features are of universal value versus those that are discriminative only for specific task pairings. This consideration has relevance to prior work in task prediction, and this consideration is discussed further in the remainder of this section.

3.2.1 Fixation Dynamics, Covert Attention and Data Partitioning

There is a clear benefit for the tasks examined in this thesis to considering fixation dynamics in conjunction with fixation densities or local image features. While viewing a scene is marked by a sequential scanpath that results from interaction between overt and covert attention mechanisms, analysis based on density alone fails to capture such temporal dynamics. This is evident in the work of Haji-Abolhassani and Clark [2014], in which accuracy in task prediction exceeds prior efforts that consider primarily static spatial characteristics [Borji and Itti, 2013; Kanan et al., 2014]. Moreover, modeling the problem in the framework of a Hidden Markov Model (HMM) allows the problem to be cast in terms of gaze points as observations, with the focus of attention as a hidden variable. This has the benefit of providing an explicit mechanism to relate overt attention to latent covert attention, and this framing of the problem also has apparent additional value in characterizing factors that distinguish different tasks.

One important property of the HMM characterization used by Haji-Abolhassani and Clark [2014], is in modeling the relationship between observed gaze locations and attended

locations. That a face is not fixated explicitly does not imply that it was not attended. This non-locality within the model allows for a richer construct for capturing the role of content not directly at the center of gaze in task prediction. In the characterization presented in this thesis, there is some non-locality to prediction in that fixation density extends outside of the centers of gaze, and image characteristics rely on features or regions that correspond to a region surrounding the center of gaze. This is more restrictive than the model of Haji-Abolhassani and Clark [2014], but does allow for some non-locality of spatial and structural information. It is also the case, that the HMM formalism considered by Haji-Abolhassani and Clark [2014] requires stronger explicit definition of stimulus characteristics such as target position, or loci of interest inferred (automatically) from fixation clusters.

The role over overlap among images in training / testing is highly relevant to the interpretation of results from prior studies that consider the Greene et al. [2012] data. In the case of spatial densities, the small image set and overlap among training and testing data implies that information on spatial position of content that is image specific may be leveraged by a classifier that considers only spatial density. For the HMM based analysis of Haji-Abolhassani and Clark [2014], this is also true as clusters of fixations, or labeled focal points within images are critical to the statistical representation in the affinity between observed fixation positions and these key locations. In contrast to general task dependent differences in center bias, or low level structure (e.g. edges) at fixation, both of these schemes present the possibility that image specific positions of high level features (e.g. objects) are a strong factor in boundaries drawn by a classifier. That is, in contrast to differences in dynamics, or general differences in spatial profile, success in delineating tasks may be relatively specific to known object positions for a known set of images.

To examine this point further, we consider differences in normalized fixation densities

across two different tasks (Memory and People). The choice of this pair is based on the relatively small degree of confusion that exists between this pairing in prior classification studies. Spatial fixation densities are produced as discussed in the methods section, and subsequently the density maps for each image are averaged within task. This implies a density that is based on an average of 4 observer's viewing patterns. Figure 3.10 shows the difference between densities associated with the Memory and People tasks, such that red regions indicate regions for which observed density is higher in the People task, and blue regions indicate those for which density is higher in the Memory task. The visualization in Figure 3.10 seems to leave little doubt that there are task dependent differences that are specific to certain types of image content. One might surmise that the failure to classify tasks above chance in the study of Greene et al. [2012] is due primarily to a lack of features that capture positions of specific relevant objects within the images considered. In contrast, alternative classification efforts [Borji and Itti, 2013; Kanan et al., 2014; Haji-Abolhassani and Clark, 2014] carry a relatively strong encoding of this type of information, even if not explicit. That is, a small image set may reveal peaks in density corresponding to items of semantic relevance, even if the analysis has no explicit knowledge of semantic categories and their relevance. Although the analysis of Greene et al. [2012] includes a measure of dwell time corresponding to specific object categories, the correspondence to discrete localized regions, and non-spatial nature of this measure may be limiting. Moreover, it is possible that measuring dwell time rather than instances of fixations might limit the discriminative value of this type of feature. As a whole, the observations concerning the importance of specific object positions suggests that models capable of identifying and localizing a large array of object types, or patterns of semantic relevance might achieve a much higher degree of success for task prediction given novel images or scenarios.

Figure 3.10: A visualization of differential fixation density (subtractive) between the *People* and *Memory* tasks. Distributions are normalized prior to calculating their difference, and may be treated as probability densities. Red regions correspond to those for which a higher density is observed with the *People* task, and blue the *Memory* task.

3.2.2 What Can Task Prediction Reveal about Vision?

In recent years, studies involving task prediction have focused principally on prediction accuracies, and on confirming Yarbus' assertions concerning task predictability. In taking a more comprehensive account of task predictability, it is important to consider what studies involving task prediction are able to convey about human vision. To this end, we discuss a few types of analysis for which task prediction may be a valuable tool while also highlighting some of the associated limitations.

1. Task Similarity: It is evident that binary classification accuracies might be used as a measuring stick for task similarity, and that the specific types of features that are discriminative for different task pairings may aid in this determination. One can imagine establishing an embedding (or topology) of task relatedness based on a large array of tasks, and suitably chosen features. However, failure to observe differences among tasks may be due to limitations in the set of features chosen. Moreover, the distances among task categories or ease of discriminating between tasks is also dependent on choosing the *right* set of features. This does not imply that that the goal of establishing a representation of *task space* should be abandoned, but does call for caution in how results are interpreted. Pushing the ceiling on accuracies achieved in task prediction will help to establish the most relevant set of features. There are also alternative sources of data, such as brain imaging data, that may provide an adjunct source of statistics for considering measures of similarity and as a useful basis for comparison to task prediction results.

2. Discriminative Features: One advantage of analysis by task prediction that is exposed in the results presented in this thesis, is the capacity to identify subtle factors that delineate different tasks. As a tool for identifying relevant features, this provides the capacity to identify relevant experimental factors that may otherwise be ignored. This also affords the potential to probe a potentially large set of features in their capacity to discriminate between tasks, and to draw out factors of greatest significance.

Chapter 4

Predicting Affective Response

4.1 Introduction

The methods applied in this chapter largely follow those previously described in Chapter. 2, with a few differences. In addition to examining the performance of the Random Forests for regression, we have also included Lasso Regression as a baseline for comparison. Determination of independent factors that explain measured affective responses are also included in the analysis, derived from Joint Approximate Diagonalization of Eigen-matrices (JADE) [Cardoso, 1994]. This also provides the capacity to determine relationships between the different measured affective response dimensions. Some of the feature sets were also modified slightly, and additional cases added to conduct analysis specific to this dataset. Given that a common feature set is used when predicting subjective ratings (affective response) one can also gain a sense of how appropriate a certain feature set is in a data dependent manner, and which gaze derived features seem to be universally valuable for predictive analysis. Unlike the task prediction data, observer data in this case was anonymous and the

distribution of viewed images was not balanced symmetrically across observers, precluding the ability to perform any analysis on the impact of partitioning across different subsets of images or observers on the correlation coefficients.

4.1.1 Subjective ratings

Lundqvist [2015] conducted a series of experiments, with each session resulting in recordings of gaze data, pupil size and 8 different subjective ratings of the stimuli. These rating scales were chosen to probe affective characteristics of the images, as rated by the viewer. During the rating of stimuli, advertisements were presented one at a time on a 30" computer screen (2560*1600 resolution), at a size corresponding to a full-page ad in a tabloid newspaper. A sample image with associated fixation densities, and subjective ratings can be found in Fig. 4.1. A larger representative sample of these images appears in Fig. 4.2. During rating, stimuli were centred vertically on the left half of the screen. To the right of the stimulus, a total of 8 rating scales were displayed in a vertical column. The scales included ratings for Brand Experience, Interest, Understanding, Relevance, Valence, Memory, Purchase Intention, and Originality, the precise description of each rating can be found in Table 4.1. When all scale values were set by the participant, a "Continue" button became visible, and the participant proceeded to the next ad by pressing the space bar. As such the amount of time spent viewing each advertisement differs. Participants navigated up and down between scales by using the up and down arrows on a keyboard, and adjusted the scales with left and right arrows. Each scale displayed a range between -1 and +1. Actual recorded values had a resolution of 300 discrete values on this scale, so scores were in the range of -150 labelled "No" to 150 labelled "Yes" for each scale. As can be seen in Fig. 4.3 the distribution for these values varies across category. We notice a common trend of high

peaks at either extreme, especially at the lower end corresponding to the "No" response. The ads were presented in random order over a session. For each trial/advertisement, both the order and the polarity of all scales were randomized.

Figure 4.1: A sample image from the Lundqvist [2015] dataset showing fixation density corresponding to aggregated observer fixations (left) and the average subjective ratings (right).

4.2 Results

Correlation coefficients for predicting user responses are found in Table. 4.2. These demonstrate a higher level of success than the Lasso Regression in Table. 4.3. The reason for

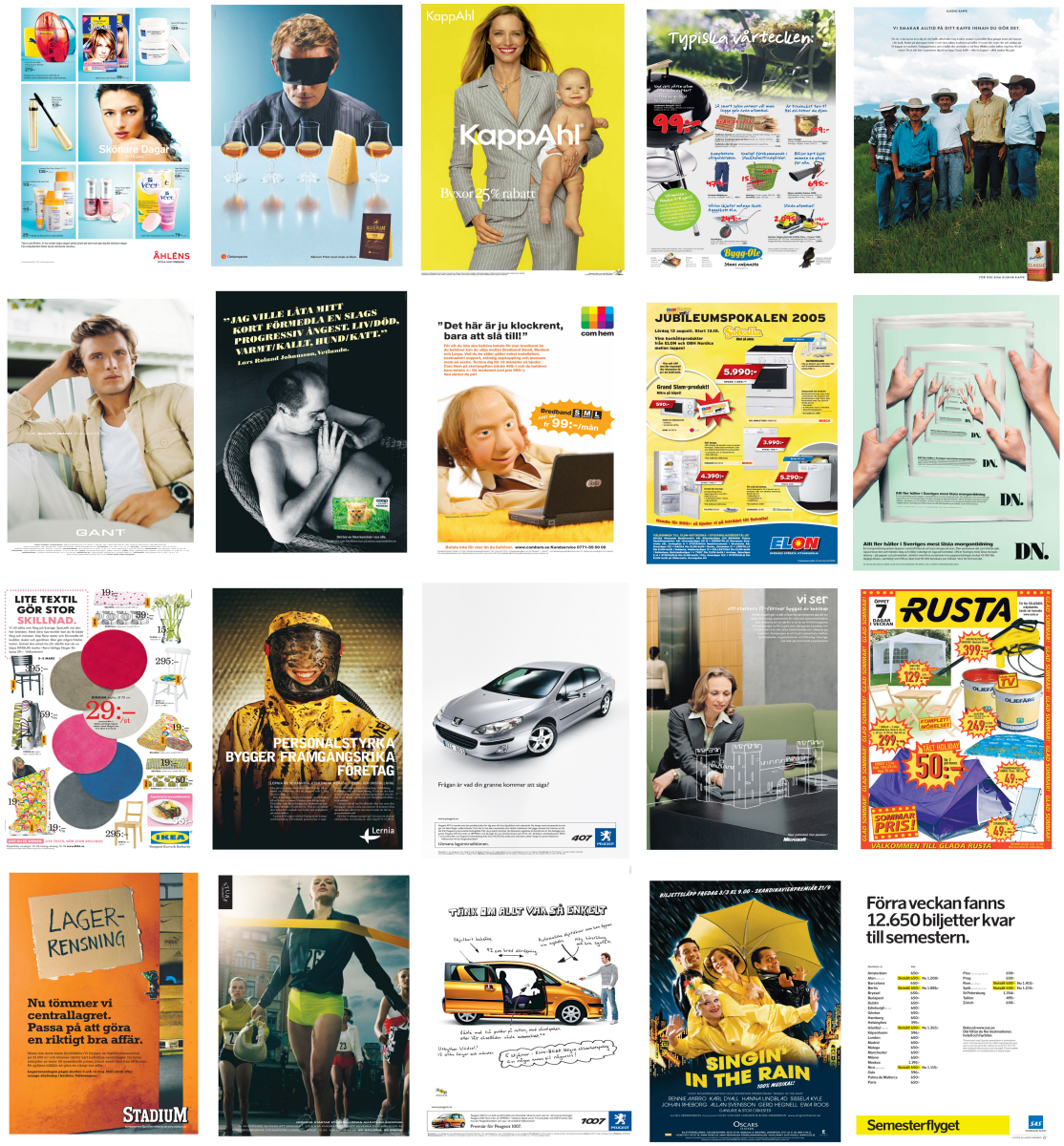

Figure 4.2: A sample of images typical of the Lundqvist [2015] dataset.

including this analysis, is to gauge the extent to which a linear fit to the measured features (gaze and image) are able to account for observer responses. It appears that the more complex classification boundary defined by Random Forests is important to successful prediction of subjective ratings in cases where this is possible. In contrast, the Lasso Regression is con-

Table 4.1: Rating Instructions given to participants, *N.B. translated from original Swedish*

Rating	Description
Brand Experience	"I have prior experience from the advertiser (have bought something, been in contact with)"
Interest	"The advertisement makes me interested"
Understanding	"The advertisement is easy to understand"
Relevance	"The advertisement directs itself/is aimed at me"
Valence	"I like the advertising
Memory	"I recognize the advertisement from previously/before this investigation"
Purchase Intention	"I want to buy the product/service that is advertised"
Originality	"The advertisement stands out from the norm/from the crowd"

strained to linear transformations of the input to predict ratings selected by participants. As such, the Lasso results may be treated as a baseline to compare the performance of Random Forests, and to assess the relative importance of inter-feature dependence. Note that this consideration also applies to individual feature types. For example, a complex boundary defined by the spatial density alone is possible using decision trees. References to performance in the next sections refer to the Random Forest results for the reasons mentioned.

4.2.1 Spatial Features

In contrast to the spatial density importance results seen in Fig. 3.4 from the task prediction analysis, those for the prediction of affective responses in Fig. 4.4 do not demonstrate the same spatial bias in any of the 8 cases, and are not strictly biased to the center. Given that the image dataset in this case is specific to ads, this might be expected. For example,

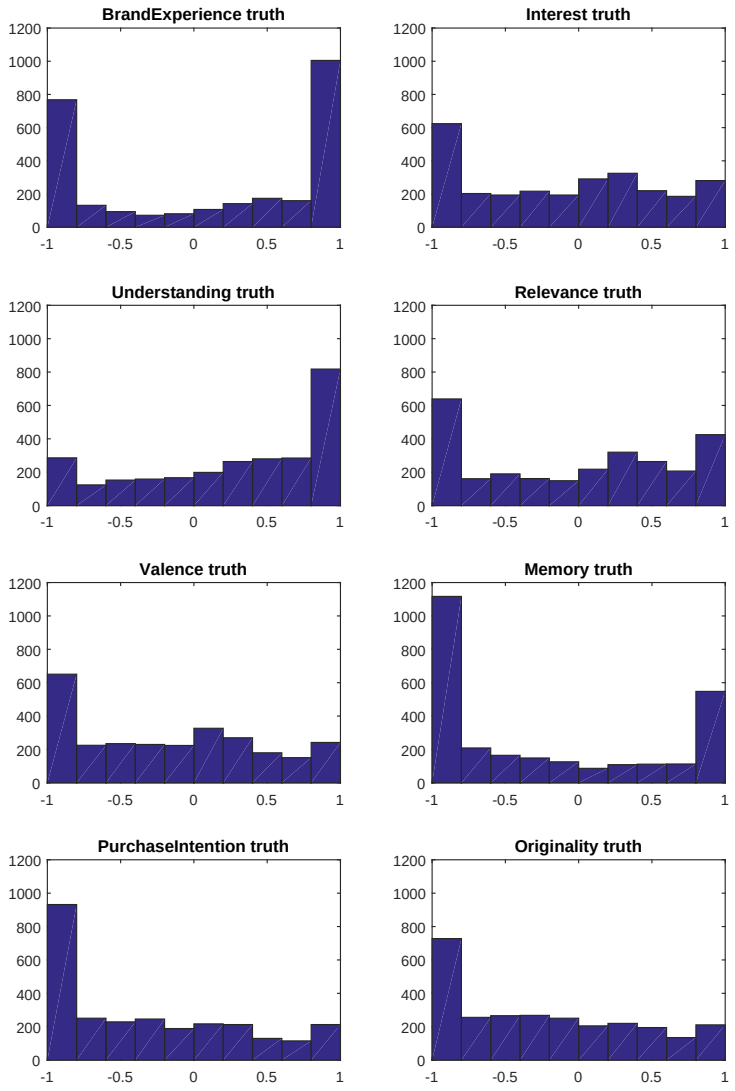

Figure 4.3: The distribution of response values for each of the subjective ratings. Note that very strong positive and negative valence is common, but there is some variation across conditions.

placement of text, logos and other common features of these images would be expected to have characteristic locations. In examining the overal fixation density, one does observe central bias Fig. 4.5, however the plots in Fig. 3.4 only indicate that this bias is not critical to discriminating between conditions.

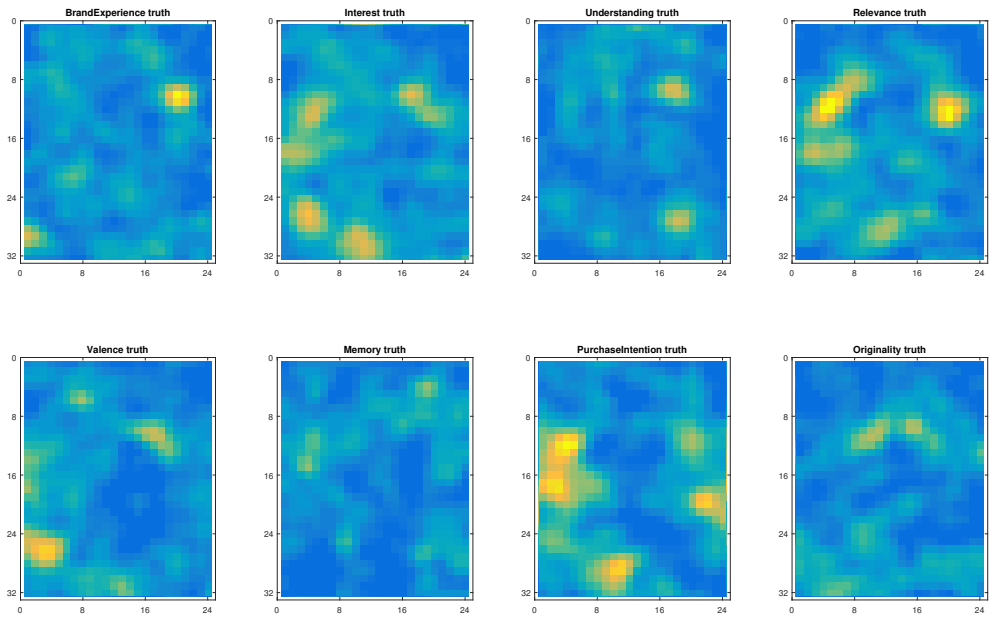

Figure 4.4: Importance of spatial density features for the subjective rating prediction problem. No clear centre bias can be observed. Note that the fixations do exhibit central bias, but this figure suggests that specific regions tend to have more diagnostic value for ratings according to the rating being assigned.

4.2.2 Local Image Features

In contrast to the task prediction problem, the results from the LM and HoG feature sets show differences in their efficacy for regression. In the case of each rating the results

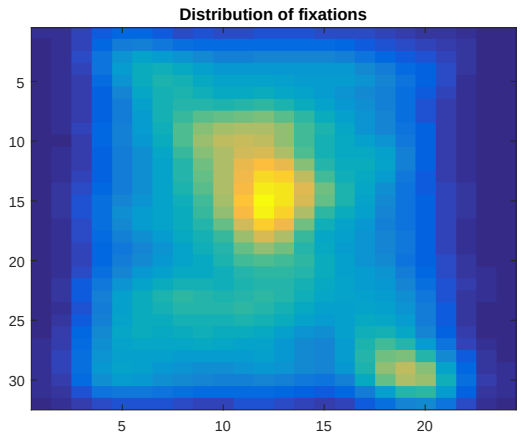

Figure 4.5: The Fixation Distribution centre bias is clear. Also note a secondary locus common for fixations to occur in the bottom right which may correspond to a common location for logos, prices or other relevant information.

from LM and HoG diverge by a minimum of a couple percentage points, in some cases being separated by as many as a dozen. This is interesting as both of these feature sets model similar image structure at fixation points. The complexity, and nature of patterns captured by LM and HoG features are different however, which may be the reason for the difference in results, especially considering both feature sets are based on mean and standard deviations of the feature vector dimensions. It is conceivable that the specific nature of patterns in the ad dataset (including many high contrast regions), in contrast to the more natural images considered for task prediction is important to the relative value of information carried by these features.

4.2.3 Quad-Nonant Features

After studying the first dataset we sought to include a feature that includes elements of both number of fixations and the spatial density. As such we have included a feature vector

composed of 14 dimensions: the total number of fixations, the number of fixations in each of the four quadrants of the images, and finally the number of fixations in the nine nonants of the image (i.e. within a 3x3 grid). This additional partitioning was an attempt to see if an alternate representation of the fixation distribution would have a positive effect on the prediction results. This appears to have benefits to the Interest and Purchase intention ratings in particular.

4.2.4 Pupil Features

Included in the Lundqvist [2015] dataset are various measures of participants pupil sizes during their observations, this subset of the dataset did not encompass all of the Image / Observer pairs so it could not be combined with the other features, and instead is examined in isolation. This feature set provides an alternative metric to gauge participants reaction to stimuli, which is also pragmatic given established relationships between pupillary size and arousal [Bradshaw, 1967; Partala and Surakka, 2003; Bradley et al., 2008; Murphy et al., 2014]. The results are comparable to other individual feature sets and there is potential for better performance in future work by combining Pupil measurements with other feature types. The datapoints composing the Pupil feature vector are: Number of Samples, Number of Valid Samples, cleanMedian, the Mean, the Standard Deviation, the Maximum Standard Deviation, the Minimum Standard Deviation, the Maximum Size, the Minimum Size. Following this are the averages of the 10 highest and 10 lowest samples in the data snippet, the average of samples at onset, 500 ms, 1000 ms, 1500 ms, and at offset. The final portion of the feature vector is composed of: the average of the difference between the average of the 10 largest and smallest values devided by the mean; the number of standard deviation between the beginning of the recording and at 500 ms, 1000 ms, and the end of recording;

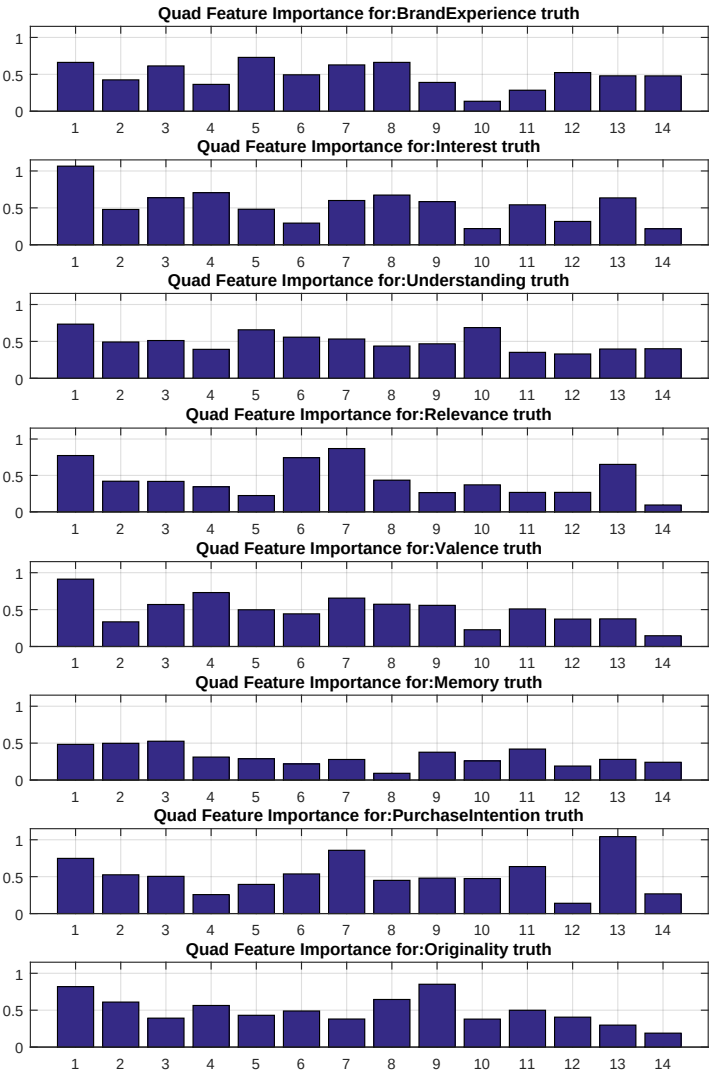

Figure 4.6: Importance measurements based on out-of-bag analysis for the Quad-Nonant feature set. This is is fairly consistent across response categories.

and the standard deviation between the maximum and minimum value.

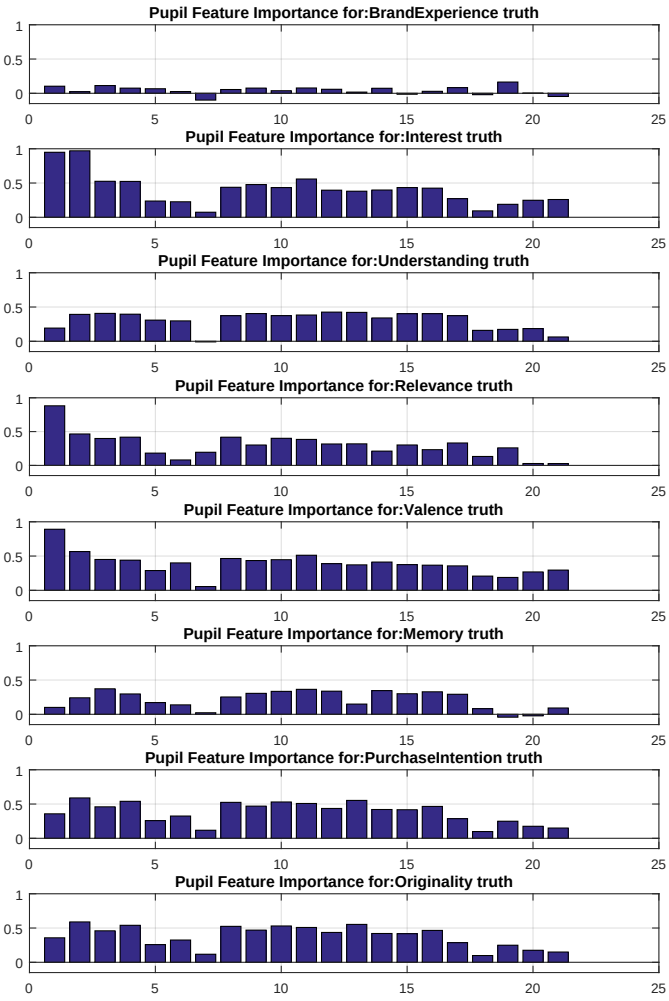

Figure 4.7: Relative importance of pupil derived statistics. Total number of samples carries a high degree of importance, other statistics have a relatively even distribution of importance, except for the measures of standard deviations and those found at the tail end of the feature set.

Table 4.2: Random Forest results for various feature combinations (Correlation)

Spatial Density	LM Filters	HoG Features	Quad-Nonant	Saccades	Num. Fixations	Gist		Brand	Interest	Understanding	Relevance	Valence	Memory	Purchase	Originality
✓								24.39**	33.84**	24.28**	28.64**	20.68**	10.47*	35.89**	22.18**
	✓							12.39**	27.49**	14.96**	17.71**	22.15**	0.88	23.4**	9.77*
		✓						25.39**	24.57**	10.75*	27.03**	14.91**	9.15*	26.09**	14.99**
			✓					23.42**	32.66**	10.48*	26.03**	14.5**	3.21	32.03**	13.87**
				✓				14.85**	36.05**	5.12	22.95**	25.9**	8.03*	31.52**	28.16**
					✓			10.78*	32.86**	1.96	23.85**	22.99**	1.55	29.46**	17.15**
						✓		19.5**	13.44**	9.65*	6.67	10.26*	12.43**	7.77	29.87**
✓	✓							25.2**	35.81**	24.45**	28.06**	22.76**	10.02*	36.41**	22.58**
✓		✓						30.05**	36.56**	24.12**	31.03**	21.66**	12.02**	37.66**	23.78**
	✓				✓			13.23**	30.13**	14.64**	22.4**	21.53**	-0.15	30.21**	11.43**
	✓		✓					18.02	31.52**	16.55**	25.14**	18.49**	0.17	31.5**	15.12**
		✓	✓					29.24**	35.39**	13.21**	32.44**	21.19**	9.09*	35**	21.05**
		✓			✓			25.69**	33.06**	10.86*	31.42**	21.8**	9.37*	33.94**	19.01**
✓						✓		29.22**	37.19**	27.28**	31.01**	23.83**	13.83**	35.95**	32.28**
✓	✓	✓	✓	✓	✓			30.44**	35.77**	24.21**	30.7**	21.51**	11.17**	38.04**	26.51**
✓	✓	✓	✓	✓	✓	✓		31.6**	37.85**	24.63**	31.46**	25.97**	14.56**	38.27**	36.32**
Pupil								2.48	30.22**	0.28	18.06**	23.76**	0.9	23.77**	18.24**

4.2.5 Saccade Features

Our Saccade feature set is composed of two distinct sections, the first half are the Saccade amplitudes manually calculated from each fixation location and are equivalent to those found in the task prediction analysis. These saccade amplitudes do not account for ocular drift that may occur between saccades. The second portion of the saccade features are saccade measurements included in the dataset. These include additional datapoints such as the start and end locations of the saccade, their amplitudes accounting for drift (produced by the

Table 4.3: Lasso Regression results for various feature combinations (Correlation)

Spatial Density	LM Filters	HoG Features	Quad-Nonant	Saccades	Num. Fixations	Gist		Brand	Interest	Understanding	Relevance	Valence	Memory	Purchase	Originality
✓								11.54**	12.56**	14.03**	16.7**	10.01*	9.84*	16.07**	8.49*
	✓							5.07	2.02	7.93	-5.97	0	-3.49	4.58	1.12
		✓						16.72**	5.43	-0.9	13.04**	0.34	2.35	14.94**	7.64
			✓					6.91	29.54**	0	19.37**	15.79**	0	28.2**	18.18**
				✓				3.1	35.04**	0	22.71**	27.14**	8.01	30.23**	25.31**
					✓			4.03	31.65**	1.66	16.22**	19.47**	-0.28	24.46**	18.17**
						✓		19.19**	5.38	6.85	0	5.54	4.81	0	30.34**
✓	✓							13.88**	29.26**	14.87**	18.9**	18.95**	7.63	25.13**	15.53**
✓		✓						17.49**	29.73**	12.18**	21.6**	18.39**	10.18*	25.56**	14.05**
	✓				✓			10.67*	31.13**	8.93	11.44*	19.31**	-3.55	23.94**	17.92**
	✓		✓					12.13**	28.2**	9.1	14.53**	16.36**	-3.31	26.55**	18.18**
		✓	✓					18.98**	29.5**	-0.79	23.05**	14.9**	2.82	28.54**	13.08**
		✓			✓			17.54**	30.68**	-0.91	21.58**	18.04**	2.74	25.94**	13.33**
✓						✓		17.4**	13.96**	13.59**	14.01**	9.85*	11.82**	16.72**	20.31**
✓	✓	✓	✓	✓	✓			17.81**	32.19**	11.56*	25.69**	23.15**	9.36*	1.25	5.88
✓	✓	✓	✓	✓	✓	✓		19.39**	33.1**	11.71*	23.41**	23.59**	11.5*	30**	27.46**
Pupil								3.16	27.09**	-2.91	17.62**	18.65**	4.72	18.62**	12.61**

eye tracker software), and its duration. To account for a varying amount of fixations and therefore feature vector length, feature vectors were padded with zeros to lengths of 125 and 800 respectively. This also implies an implicit encoding of the number of fixations within the feature vector.

In considering feature importance, the two different aforementioned sets of saccades statistics were examined separately. Analysis of feature importance reveals higher importance in the first datapoints which gradually decreases with each successive fixation. The

saccade features performed best when predicting Interest, and Valence, outperforming all other individual feature sets. Saccades were also second to Gist in predicting Originality, at a level which few other sets approached. It is interesting to note that with ascending index in the feature vector, negative weights begin to appear suggesting no value beyond a certain period of viewing.

4.2.6 Global Image Features

For task prediction, the Gist features were of little value alone in determining task. This is not surprising since the same images were viewed across the different tasks. In contrast, the subjective ratings selected are associated directly with the image content. In this case, Gist outperformed all other feature sets when predicting Originality, with saccades as the only comparable feature set. While the overall performance of all feature sets when considering Memory was comparatively low to other responses, Gist stood as the single best predictor. In the section that follows, results also indicate that Gist combines well with other features, in improving predictions.

4.2.7 Feature Combinations

It is immediately clear that combining features is beneficial to the correlation coefficients, both Brand Experience and Originality exceed the 30% mark. While Memory remains low, it benefits somewhat from the increased number of features. The inclusion of Gist with the other feature sets has a strong impact on the results, particularly in the case of Originality. This is especially interesting if we remember that Saccades had a similar level of performance when comparing the individual feature sets but that combination of saccades with the other features did not demonstrate an improvement in Originality performance. As we can see

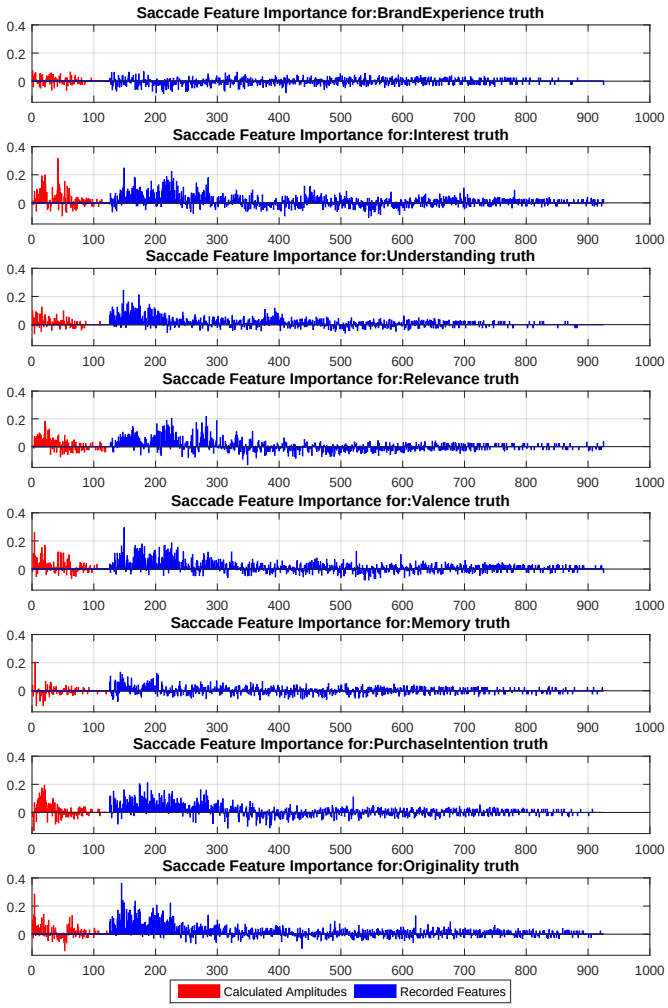

Figure 4.8: Relative Importance of the Saccade features, those in red denoting the amplitudes manually calculated with the fixation locations and those in blue denoting the data-points generated by the eye-tracker. To compensate for varying number of fixations each vector was padded with zeros, likely affecting the importance of later features.

in Fig. 4.8 each rating shows differential benefit from different types of features, while the combinations of all the features ensures a level of performance that approaches the optimum.

Some important general observations may be made regarding feature combinations:

1. The Spatial Density feature set, while relatively large compared to the others, has a consistent level of performance and performs well in combination with other feature sets.

2. The number of fixations and quad-nonant features give good performance and augment performance when combined with other feature sets suggesting independence in the information they carry.

3. The Gist image features are very capable in prediction performance but also benefit greatly from combination with other features. This is critical as it indicates that the nature of the advertisement is important to driving affective responses, but associated behavioural measurements help to draw out individual differences.

4. In this particular case, the LM and HoG filters did not greatly outperform the Spatial features in any particular case. They do achieve similar results in many instances with a much smaller feature set, their combination with the Spatial features also shows a noticeable improvement in performance.

5. The Saccade features produce respectable prediction performance, similar to the spatial features in many cases. These are especially important to the prediction of Originality. Unlike the Spatial features however, there is less value for Interest, favouring Originality and Valence rather than Relevance and Purchase Intention. This begins to hint at correlation between subjective measurements, which is examined in the section that

follows in the context of the Jade algorithm.

4.2.8 JADE

As discussed in Chapter 2, we attempt to reduce the subjective ratings to a set of independent feature dimensions based on independent component analysis (ICA) using JADE. Examining the results of JADE reveals some interesting insight. Fig. 4.9 shows a visual depiction of the relative weight assigned to individual raw measured subjective ratings that form the assumed independent sources that combine to produce the measured subjective ratings. Two of the rows exhibit a singular strong weight for Memory and Brand Experience respectively. In contrast, the other rows present a more widespread distribution of weights shared across subjective ratings. In particular, Interest, Relevance, Valence, and Purchase Intention are strongly related in the axes they are pooled upon. The combination of these 4 particular subjective ratings can be explained by the high correlation observed between them in Fig. 4.10. The lack of correlation between Memory and the other ratings also explains it's representation in the Jade components. It is also notable that the 4 correlated measures are among those that tend to be easiest to predict on average across different types of features.

Also of note is the 3rd Jade component, which features a strong weight for the Memory question. This component presents an overall poor prediction performance, which is not surprising given the low coefficients for Memory predictions based on the raw measurements. In contrast, the 5th component has high weights on Interest, Relevance, Valence, and Purchase Intention, which feature some of the highest coefficients, and the associated JADE component is also revealed to be highly predictable. It is also interesting to note that while the 4th component has a strong negative weight for Brand Experience, prediction of the value of this component tends to be higher than for Brand Experience in it's raw form, sometimes

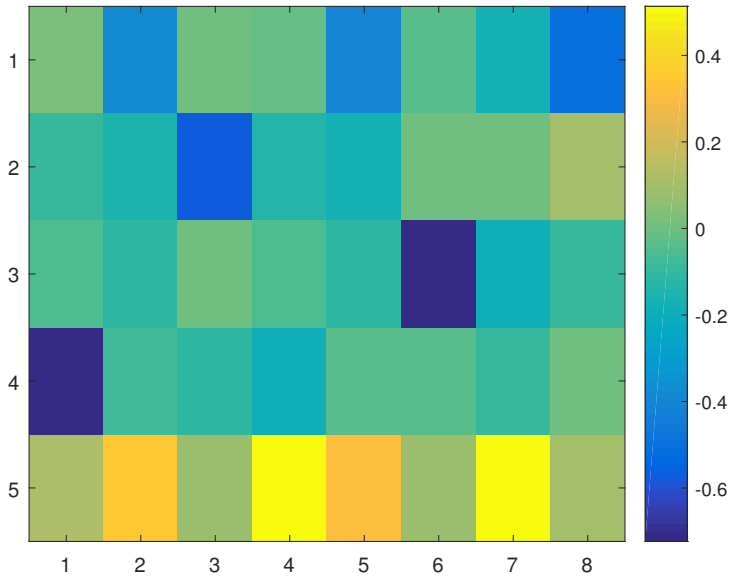

Figure 4.9: The relative weights assigned to the original 8 ratings when translated into the associated 5 JADE independent components. Note that weights indicate the independent sources that are assumed to combine linearly to produce measured subjective ratings. The reason there are 5 components, is that PCA as a pre-processing step reveals that this captures the vast majority ($> 98\%$) of variance in the raw measurements.

by a significant amount (e.g. based on saccades and LM Filters). Another surprising result is that of the 2nd JADE component, which has a strong negative weight in Understanding and a small positive weight for Originality, which also shows good capability for prediction. As a whole, these results present some interesting insight into the data, and also the potential for more forensic analysis of the meaning of these features with respect to their weights relative to the original measurements. For example, the opponency between understanding and originality for the 2nd component is an interesting case for careful examination at the

level of how this correlates with actual image patterns.

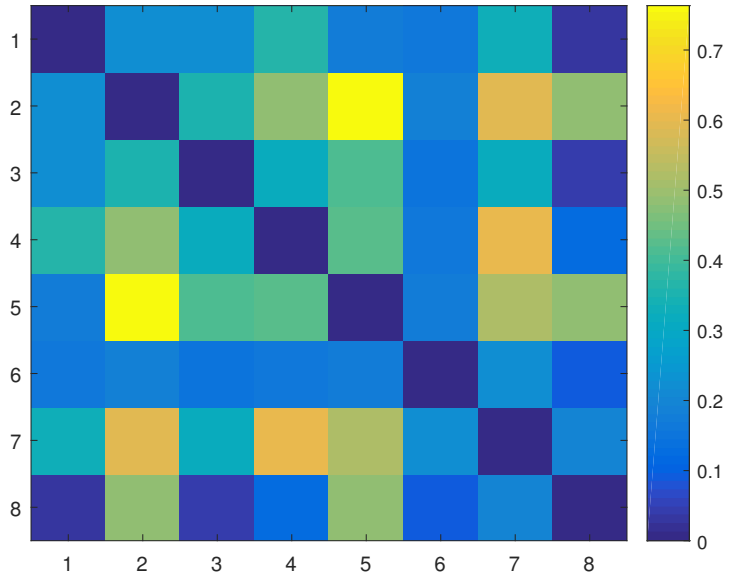

Figure 4.10: Correlations between all 8 each subjective rating. Ratings 2, 4, 5, and 7, corresponding to Interest, Relevance, Valence, and Purchase Intention, all share relatively high correlation coefficients between them. We also notice that Rating 6, Memory, is the only one with no coefficients above 0.25. BrandExperience is similar although it does have slightly higher correlation with Relevance and Purchase Intention.

4.2.9 T-SNE

As another means of visualizing the relationship between ratings we used the T-Stochastic Neighbour Embedding technique for dimensionality reduction (T-SNE) algorithm from Van der Maaten and Hinton [2008]. This provides an embedding of the raw subjective ratings such that their position in the embedding is as close as possible to their position in the raw 8-

Table 4.4: Random Forest Jade results (Pearson Correlation)

Spatial Density	LM Filters	HoG Features	Quad-Nonant	Saccades	Num. Fixations	Gist	Jade 1	Jade 2	Jade 3	Jade 4	Jade 5
✓							32.07**	29.38**	22.34**	31.17**	37.18**
	✓						12.84**	11.35**	2	9.94*	23.62**
		✓					13.56**	7.06	7.96	23.57**	22.41**
			✓				27.27**	17.07**	11.2**	26.36**	33.78**
				✓			35.86**	16.31**	17.52**	17.18**	33.69**
					✓		13.32**	7.2	8.79*	23.97**	22.14**
						✓	30.16**	10.5*	12.62**	20.53**	13.89**
✓	✓						26.32**	20.54**	9.85*	21.67**	33.6**
✓		✓					24.87**	18.42	10.11*	26.79**	34.56**
	✓				✓		27.92**	22.33**	8.08*	19.82	32.28**
	✓		✓				33.06**	14.46**	21.45**	-0.27	11.73**
		✓	✓				34.98**	27.36**	25.73**	7.73	9.29*
		✓			✓		31.65**	24.07**	16.83**	30.03**	37.85**
✓						✓	36.41**	24.83**	13.58**	27.91**	34.84**
✓	✓	✓	✓	✓	✓		27.69**	19.11**	9.94*	27.71**	33.97**
✓	✓	✓	✓	✓	✓	✓	33.31**	32.52**	21.32**	33.71**	37.84**
Pupil							2.2	30.25**	0.06	18.07**	24.06**

dimensional space. As can be seen in Fig. 4.11 this produces clusters of ratings, that we have labelled for analysis.

Looking at the graph in Fig. 4.12 we notice that Group 1 systematically contains only negative ratings, distinguishing it from all other groups. Further analysis showed that this group is composed of observations from the majority of users, alleviating concerns of this being caused by a minority, and overly critical, observers. Group 2 contains high ratings in both Brand Experience and Memory, as well as a high concentration of other positive

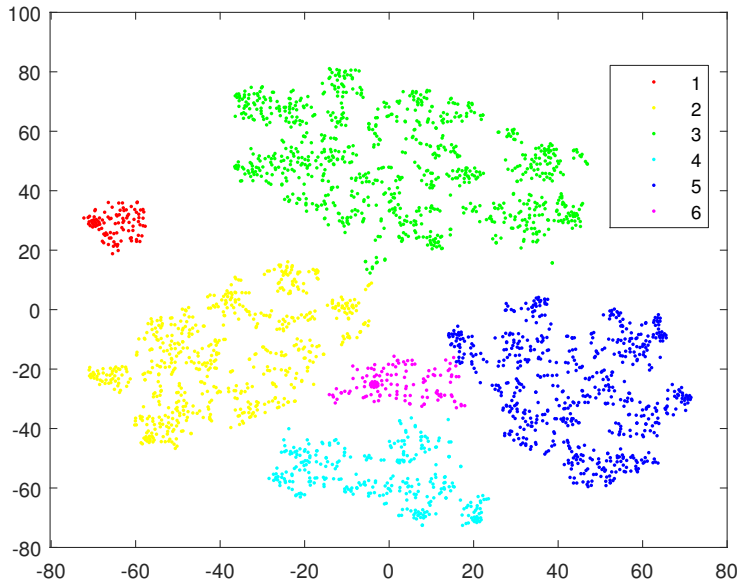

Figure 4.11: Each observation's 8 subjective ratings were embedded into a 2-dimensional mapping based on t-Stochastic Neighbour Embedding (T-SNE). This results in a topology that seems to include 6 clusters, these were labelled to facilitate viewing and discussion

values originating from it's bottom-left corner. Group 3 contains high ratings for Brand Experience but low Memory ratings, it is also interesting to note the separation of high Purchase Intention and Originality ratings along the center axis of the left side while Interest remains high in the same region. Group 4 contains low ratings for Brand Experience and high Memory ratings. It is an otherwise diverse group in regard to the other ratings with higher values tending to the left side. Group 5 contains low ratings for both Brand Experience and Memory, higher ratings of the other categories are found primarily towards the bottom right corner of the group. Group 6 contains neutral ratings approaching 0, it is befitting that it found in the approximate centre of groups 2 to 5.

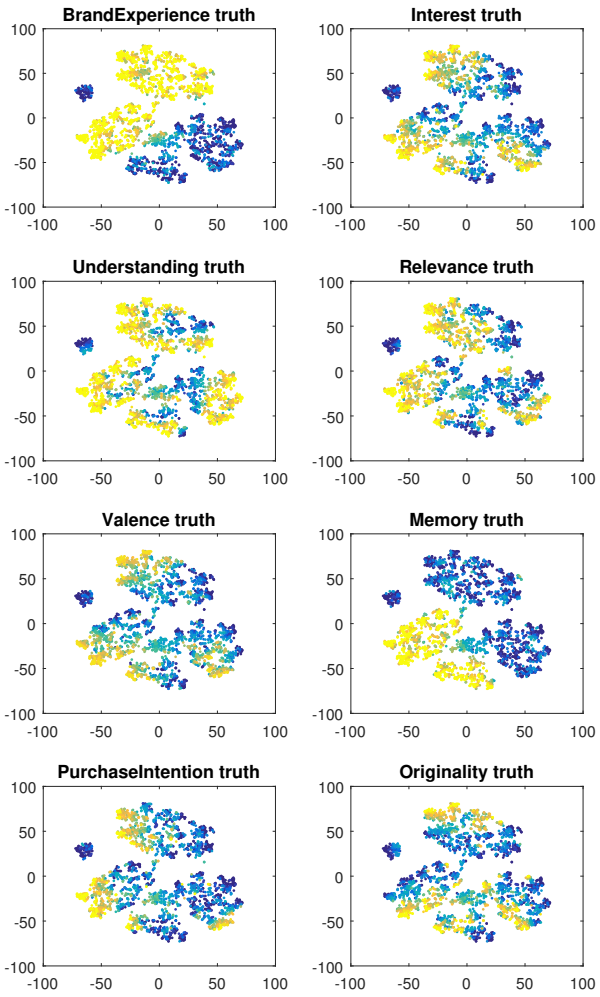

Figure 4.12: Heatmap of subjective responses across the 6 T-SNE clusters. Group 1 contains only low ratings across all subjective measures; Group 2: high Memory and high Brand Experience; Group 3: low Memory and high Brand Experience; Group 4: high Memory and low Brand Experience; Group 5: low Memory and low Brand Experience; and finally Group 6 contains ratings closely centred around 0.

From these observations it is clear that the separation performed by T-SNE revolves around the values of both Brand Experience and Memory, which we have previously observed as being the two ratings with the least amount of correlation with the others. This provides an additional visualization of the participant data in regards to the relative prevalence of different types of ratings, and the bias and correlation expressed across different ratings.

4.3 Discussion

4.3.1 Different Features for Different Responses

We have considered the extent to which the affective responses of participants for visual stimuli can be predicted by the spatial density profile of fixations, features at fixations, scene structure, fixation dynamics, or pupil measurements. In addition to determining the type of features most useful in inferring each aspect of a participant's response, we have also observed the benefits in considering the correlation structure of the response ratings, both in analysis and visualization.

Important observations from this analysis are summarized in what follows:

1. The Brand Experience and Memory Tasks are most disjoint from the other ratings. In contrast, others showed some level of correlation, especially in the case of Interest, Relevance, Valence, and Purchase Intention.

2. The different response ratings benefited from different feature sets to achieve the best performance, and the combinations of all features was found to be especially capable for most cases. One case that stands out is that of Gist which contributed a great amount to the Originality results, but performed relatively poorly for other ratings.

3. The Saccade Amplitudes performed well for the highly correlated ratings of Interest, and Valence, as well as the moderately correlated Originality. This is particularly interesting as other features sets with good performance for Interest and Valence did not perform as well on Originality but performed better on Relevance and Purchase Intention which, while highly correlated with Interest and moderately with Valence, are poorly correlated with Originality.

4. Spatial density, HOG, and Quad-Nonant features showed good results overall and in particular when inferring Brand Experience, Interest, Relevance, and Purchase Intention. The Spatial features also greatly outperformed the others when inferring Understanding.

5. In comparison to the other ratings, Memory could not be inferred above a level of 12% from Gist and 10% from Space with the combined features achieving 14.5%. It is possible that a feature set that was not examined in this thesis could yield better results, this may of course be the case for the other ratings as well. There was a very interesting improvement in the 3rd JADE component which was overwhelmingly weighted towards Memory, up to 25%. This also highlights a particular relationship between raw measurements that would benefit from detailed examination.

4.3.2 What Can Predicting Affective Response Reveal about Vision?

The analysis of the features most successful in prediction, correlation between the ratings, and determination of independent sources is an important initial step in elucidating the underlying neural or behavioural mechanisms associated with different subjective judgments.

The results presented in this chapter, and associated methodology carries some interesting implications:

1. Considering the relationship between subjective ratings and their correlation structure is relatively important to analysis of the data. Moreover, this analysis may feed back into pruning the set of questions presented in gathering the human data, independently, or in concert with characteristics that can be predicted from human behaviour produced from additional experimentation.

2. The analysis shows a surprising degree of diversity in the types of features that are of greatest value in predicting different types of subjective ratings. This provides some insight into the relationship between observed behaviours, and affective dimensions of stimuli and also suggests interesting avenues for human experiments.

3. From an application perspective, results indicate that many affective characteristics of an image may be predicted from the image content alone. However, coupled with some rudimentary human data, one can gain even stronger insight into the media in question. These considerations might be applied in successive stages to prune an initial set of candidate ads, posters or other forms of media based on appearance, and subsequent pruning with a small set of human trials to identify the most likely exemplars to be effective for advertising, web design or other application domains.

Chapter 5

Conclusion

In this thesis, we have conducted a number of tests to determine the extent to which an observer's Task (Chap. 3) or Affective Responses (Chap. 4) may be determined based on image characteristics, and/or gaze statistics. This has been carried out with a focus on the value of different factors such as spatial distribution of fixations, the nature of content at fixated locations, fixation dynamics, or overall image characteristics. This analysis is accompanied by discussion of methodological considerations important to interpreting results derived from task or subjective rating prediction. This provides a number of important observations relevant to the specific tasks considered as well as prior studies involving task prediction, and to methods important for future work involving task prediction:

1. Unlike alternative studies involving task classification, we include a heavy emphasis on feature importance. This provides an indication of the relative importance of spatial fixation density, local image structure, holistic scene structure and fixation dynamics in distinguishing between different tasks and affective responses. We observe that spatial density, and timing and length of saccades are important factors for classification, while

global image features are imperative for certain regression cases involving affective response to imagery. We also present evidence that for finer task distinctions, specific information about spatial positions of important objects may be relatively important to defining task differences. The success of some prior work may be owed to an implicit representation of the position of these targets carried by fixation density that has implications for the generality of the approach used, and how data is divided.

2. While previous work has included measures of fixation density, saccade statistics and image salience, the current study marks the first effort to examine the value of local image content at fixated locations in predicting task for a Yarbus style paradigm. Results indicate that fixated content is of significant value in delineating some tasks, even if less important than other factors. While one might expect some bias in fixated structure due to differences in spatial profile, the combined strength of spatial fixation densities and structure of content at fixation indicates that fixated features carry diagnostic information that is independent of spatial position. This impact is evident only when local image features and task prediction pairings are separated and analysed.

3. Analysis built on decomposition into different combinations of features also gives rise to a better understanding of what information can be derived from which feature sets, and which different types of features have overlap or redundancy in what they represent. In some instances, the combined strength of global image properties and gaze derived statistics are more valuable than these individual components.

4. The knowledge of which features perform best for a given task or response category can also help to determine a much smaller subset of the most critical features. This is useful from a processing or memory standpoint, but also may be of value in identifying

important targets for human experimentation, or determining an optimal candidate subset of images according to some specific objective.

5. We have demonstrated that the means of partitioning experimental data is a very important factor in interpreting outcomes from task prediction. In particular, the value of spatial densities may be relatively specific to the content of known images for some task pairings, and less so for others. This also presents the possibility that success in some prior task prediction studies may be due to implicit representation of relatively high-level contextual, or object specific factors within the spatial density associated with a smaller set of *known* images.

6. The task prediction analysis presented in this thesis is distinguished from prior efforts in the size of the data sets, the task definitions considered, and methods that are used. Differences in gaze behaviour depend on both task definitions, and image content. It is our hope that as further research efforts in this domain continue to cover a more diverse range of task/data combinations, that stronger conclusions may be drawn concerning task relatedness and underlying mechanisms.

7. Our work demonstrates that more simplistic linear regression on image and gaze statistics is much less successful in predicting user behaviour, which points to the interdependence of features as carrying important information. More robust models such as Random Forests we have employed offer a means to effectively leverage the strengths of different feature sets to achieve versatile performance, while also providing tools for measuring relative feature importance.

8. The use of blind source separation (JADE) has been demonstrated to provide insightful analysis, and to yield a set of feature dimensions wherein poor prediction performance

for individual features is improved significantly due to category-opponency that is introduced (e.g. subjective ratings combined with different signs). This analysis also allows a means to understand which ratings may derive from similar cognitive mechanisms.

9. Strategies that are successful in task and affective response prediction are also of value in application domains that include human-machine interaction, perceptual user interfaces and assistive technology for physiological or neurological conditions. This thesis contributes to the growing body of strategies for task prediction towards supporting this goal. In particular, models that include a strong representation of both dynamics and recognition of patterns with semantic relevance (e.g. objects) may be expected to be especially capable for these types of applications. The recent success of deep neural networks in computer vision will help in allowing stronger semantic information to be leveraged by predictive systems of this nature.

As a whole, this work contributes to the growing body of efforts that support Yarbus' assertions concerning fixation patterns. On the basis of the results presented, we are also optimistic that future efforts that emphasize task relevance will be fruitful in understanding task-gaze interaction. Further proliferation of efforts of this variety may also provide a window into the *bigger picture* of generalized task relatedness, task-data interaction and individual and general differences in gaze-task relationships.

Bibliography

A. Borji and L. Itti. Defending Yarbus: Eye movements reveal observers' task. Soon to be published in Journal of Vision, 2013.

A. Borji and L. Itti. Defending yarbus: Eye movements reveal observers' task. *Journal of vision*, 14(3):29, 2014.

M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607, 2008.

J. Bradshaw. Pupil size as a measure of arousal during information processing. 1967.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

N. Bruce and J. Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2006.

N. D. Bruce. Towards fine-grained fixation analysis: distilling out context dependence. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 99–102. ACM, 2014.

A. Bulling, C. Weichel, and H. Gellersen. Eyecontext: Recognition of high-level contextual

cues from human visual behaviour. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 305–308. ACM, 2013.

G. T. Buswell. *How people look at pictures*. University of Chicago Press Chicago, 1935.

J.-F. Cardoso. On the performance of orthogonal source separation algorithms. In *Proc. EUSIPCO*, pages 776–779, Edinburgh, Sept. 1994.

J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, Dec. 1993.

M. S. Castelhano, M. L. Mack, and J. M. Henderson. Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3), 2009.

M. Cerf, J. Harel, A. Huth, W. Einhäuser, and C. Koch. Decoding what people see from where they look: Predicting visual stimuli from scanpaths. In *Attention in Cognitive Systems*, pages 15–26. Springer, 2009.

X. Chen and G. J. Zelinsky. Real-world visual search is dominated by top-down guidance. *Vision research*, 46(24):4118–4133, 2006.

M. I. Coco and F. Keller. Classification of visual and linguistic tasks using eye-movement features. *Journal of vision*, 14(3):11, 2014.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

M. DeAngelus and J. B. Pelz. Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, 17(6-7):790–811, 2009.

P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *ICML*, volume 96, pages 148–156, 1996.

M. R. Greene, T. Liu, and J. M. Wolfe. Reconsidering Yarbus: A failure to predict observers task from eye movement patterns. *Vision research*, 62:1–8, 2012.

A. Haji-Abolhassani and J. J. Clark. Realization of an inverse yarbus process via hidden markov models for visual-task inference. *Journal of Vision*, 11(11):218–218, 2011a.

A. Haji-Abolhassani and J. J. Clark. Visual task inference using hidden markov models. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011b.

A. Haji-Abolhassani and J. J. Clark. An inverse yarbus process: Predicting observers task from eye movement patterns. *Vision research*, 103:127–142, 2014.

M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194, 2005.

J. M. Henderson, S. V. Shinkareva, J. Wang, S. G. Luke, and J. Olejarczyk. Predicting cognitive state from eye movements. *PloS one*, 8(5):e64937, 2013.

W. Jones and A. Klin. Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism. *Nature*, 504:427–431, 2013.

C. Kanan, N. A. Ray, D. N. Bseiso, J. H. Hsiao, and G. W. Cottrell. Predicting an observer's task using multi-fixation pattern analysis. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 287–290. ACM, 2014.

K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein. What do saliency models predict? *Journal of vision*, 14(3):14, 2014.

K. Kunze, Y. Utsumi, Y. Shiga, K. Kise, and A. Bulling. I know what you are reading: recognition of document types using mobile eye tracking. In *Proceedings of the 17th annual international symposium on International symposium on wearable computers*, pages 113–116. ACM, 2013.

T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.

D. Lundqvist. Lundqvist advertisement dataset, July 2015. URL `ki.se`.

M. Mills, A. Hollingworth, S. Van der Stigchel, L. Hoffman, and M. D. Dodd. Examining the influence of task set on eye movements and fixations. *Journal of vision*, 11(8):17, 2011.

P. R. Murphy, J. Vandekerckhove, and S. Nieuwenhuis. Pupil-linked arousal determines variability in perceptual decision making. 2014.

T. O'Connell and D. Walther. Fixation patterns predict scene category. *Journal of Vision*, 12(9):801, 2012. doi: 10.1167/12.9.801.

A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

T. Partala and V. Surakka. Pupil size variation as an indication of affective processing. *International journal of human-computer studies*, 59(1):185–198, 2003.

C. A. Rothkopf, D. H. Ballard, and M. M. Hayhoe. Task and context determine where you look. *Journal of vision*, 7(14):16, 2007.

Y. Sugano, H. Kasai, K. Ogaki, and Y. Sato. Image preference estimation from eye movements with a data-driven approach. *Journal of vision*, 7(3):1–9, 2014.

B. W. Tatler. 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):4, 2007.

B. W. Tatler and B. T. Vincent. The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17(6-7):1029–1054, 2009.

B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision research*, 45(5):643–659, 2005.

B. W. Tatler, R. J. Baddeley, and B. T. Vincent. The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision research*, 46(12):1857–1862, 2006.

B. W. Tatler, N. J. Wade, H. Kwan, J. M. Findlay, and B. M. Velichkovsky. Yarbus, eye movements, and vision. *i-Perception*, 1(1):7, 2010.

B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5):5, 2011.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

P.-H. Tseng, I. G. Cameron, G. Pari, J. N. Reynolds, D. P. Munoz, and L. Itti. High-throughput classification of clinical populations from natural viewing eye movements. *Journal of neurology*, 260(1):275–284, 2013.

L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

T. P. Vogl, J. Mangis, A. Rigler, W. Zink, and D. Alkon. Accelerating the convergence of the back-propagation method. *Biological cybernetics*, 59(4-5):257–263, 1988.

J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

H. Yang and G. J. Zelinsky. Visual search is guided to categorically-defined targets. *Vision research*, 49(16):2095–2103, 2009.

A. L. Yarbus. *Eye movements and vision*. New York: Plenum, 1967.