

Decoding attentional load in visual perception: a signal processing approach

Luke Palmer

Institute of Cognitive Neuroscience

University College London



Under the supervision of Prof. Nilli Lavie and Dr. Sam Schwarzkopf

Submitted for the degree of PhD, September 2016

Abstract

Previous research has established that visual perception tasks high in attentional load (or 'perceptual load', defined operationally to include either a larger number of items or a greater perceptual processing demand) result in reduced perceptual sensitivity and cortical response for visual stimuli outside the focus of attention. However, there are three challenges facing the load theory of attention today. The first is to describe a neural mechanism by which load-induced perceptual deficits are explained; the second is to clarify the concept of perceptual load and develop a method for estimating the load induced by a visual task a priori, without recourse to measures of secondary perceptual effects; and the third is to extend the study of attentional load to natural, real-world, visual tasks. In this thesis we employ signal processing and machine learning approaches to address these challenges. In Chapters 3 and 4 it is shown that high perceptual load degrades the perception of orientation by modulating the tuning curves of neural populations in early visual cortex. The combination of tuning curve modulations reported is unique to perceptual load, inducing broadened tuning as well as reductions in tuning amplitude and overall neural activity, and so provides a novel low-level mechanism for behaviourally relevant failures of vision such as inattention blindness. In Chapter 5, a predictive model of perceptual load during the task of driving is produced. The high variation in perceptual demands during real-world driving allow the construction of a direct fine-scale mapping between high-resolution natural imagery, captured from a driver's point-of-view, and induced perceptual load. The model therefore constitutes the first system able to produce a priori estimates of load directly from visual characteristics of a natural task, extending research into the antecedents of perceptual load beyond the realm of austere laboratory displays. Taken together, the findings of this thesis represent major theoretical advances into both the causes and effects of high perceptual load.

Table of Contents

Abstract.....	2
List of Figures and Tables.....	5
Acknowledgements.....	8
Preface	9
1 General introduction	13
1.1 Visual attention and perceptual load theory.....	13
1.1.1 Behavioural evidence for effects of perceptual load.....	17
1.1.2 Neuroimaging evidence for effects of perceptual load.....	21
1.2 Modulation of feature-specific representation	24
1.3 The causes of perceptual load.....	28
1.4 Perceptual load in driving.....	34
2 Methodological background.....	39
2.1 Measuring and interpreting brain activity	39
2.1.1 Measuring activity.....	39
2.1.2 fMRI data analysis.....	43
2.2 Computer vision and machine learning.....	54
2.2.1 Obtaining ground-truth perceptual load values.....	55
2.2.2 Extracting semantics from imagery	56
2.2.3 Regressing from descriptors to attributes.....	67
2.2.4 Tuning model hyperparameters	71
3 The effect of perceptual load on representations of orientation	73
3.1 Chapter Introduction	73
3.2 Experiment 1.....	75
3.2.1 Methods	76
3.2.2 Results	79
3.2.3 Discussion.....	80
3.3 Experiment 2.....	81
3.3.1 Methods	81
3.3.2 Data analysis.....	86
3.3.3 Results	90
3.4 Chapter Discussion.....	95
4 The effect of perceptual load on population coding of orientation.....	99
4.1 Chapter Introduction	99
4.2 Methods.....	102
4.3 Data Analysis.....	105

4.4	Results.....	109
4.4.1	Behavioural results.....	109
4.4.2	fMRI results.....	111
4.5	Chapter Discussion.....	118
5	Modelling perceptual load in driving.....	122
5.1	Chapter Introduction.....	122
5.2	Methods.....	125
5.2.1	Building a video dataset.....	125
5.2.2	Estimating ground-truth values of perceptual load.....	130
5.3	Results 1: Ground-truth perceptual load values.....	132
5.4	Data analysis.....	134
5.4.1	Video representation.....	134
5.4.2	Regression and model fitting.....	137
5.5	Results 2: Predicting load.....	140
5.5.1	Original IDT and C3D configurations.....	140
5.5.2	IDT+C3D with linear kernel.....	144
5.5.3	IDT+C3D with nonlinear kernels.....	147
5.5.4	IDT+C3D results summary.....	155
5.6	Chapter Discussion.....	156
6	General discussion.....	159
6.1	Summary of findings.....	159
6.2	Perceptual load and orientation encoding.....	162
6.2.1	The unique effect of perceptual load on orientation tuning.....	162
6.2.2	Perceptual consequences of load-induced modulations.....	164
6.3	Modelling load.....	167
6.3.1	Relation to other models of perceptual load.....	167
6.3.2	Contributions to computer vision.....	170
6.3.3	Applied implications.....	172
6.4	Concluding remarks.....	174
	References.....	175

List of Figures and Tables

Figures

Figure 1-1. Example displays used in the three conditions of Lavie and DeFockert (2003).....	15
Figure 1-2. Perceptual load manipulation in a response competition design	18
Figure 1-3 An illustrative example of a low-load display used by Forster and Lavie (2008) containing a completely irrelevant distractor; a cartoon character.....	18
Figure 1-4. Psychophysical tuning modulation due to perceptual load.	27
Figure 1-5. Example displays from Experiment 1 by Torralbo and Beck (2008).	30
Figure 1-6. Displays used by Roper et al. (2013).	32
Figure 1-7. Predicted perceptual load induced response-competition task displays.....	33
Figure 1-8. Screenshots of driving simulator in the experimental conditions employed by Marciano and Yeshurun (2015).	35
Figure 1-9. Example of low perceptual load in Murphy and Greene's (2016) simulator experiment. This represents a critical trial, as a pedestrian is placed at the side of the road. ...	36
Figure 2-1. BOLD signal in an image matrix.	42
Figure 2-2. The haemodynamic response (that measured using fMRI) induced by a short burst of neural activity (from Heeger & Rees, 2002).	43
Figure 2-3. Example GLM configuration	45
Figure 2-4. Estimation of population receptive fields (pRFs).	47
Figure 2-5. Delineation of visual areas.	48
Figure 2-6. Putative responses of two voxels to two experimental conditions.	49
Figure 2-7. Prediction of stimulus orientation with MVPA.	51
Figure 2-8. Voxel tuning functions obtained when participants were told to attend to either the 45° (black curves) or 135° (blue curves) in an overlapping grating stimulus.	52
Figure 2-9. Sample video frames with associated action categories from the UFC-101 (top row) and Hollywood2 (bottom row) action recognition datasets.	58
Figure 2-10. Dense trajectories extraction.	60
Figure 2-11. 1-dimensional representation of CNN hierarchy and local connectivity.	63
Figure 2-12. Simplified LeNet-5 (LeCun, 1998) CNN architecture set up to recognise an object in an image.	64
Figure 2-13. 2D and 3D convolution for images and video.	65
Figure 2-14. C3D architecture.	65
Figure 3-1. The RSVP task used to manipulate load in the experiment.....	77
Figure 3-2. Orientation change detection accuracy.	80
Figure 3-3. Schematic of Experiment 2.....	84
Figure 3-4. Depiction of the retinotopic mapping stimuli.	85
Figure 3-5. Mean GLM parameter values for visually responsive voxels	91
Figure 3-6. Classification performance for voxels in V1.	93
Figure 3-7. Classification results for voxels in V2 (left) and V3 (right).	95
Figure 4-1. Multiplicative (left) and bandwidth (right) feature-dependent scaling of orientation tuning curves (adapted from Liu and Carrasco, 2009)	102
Figure 4-2. Schematic of an orientation discrimination trial during the experiment.....	104

Figure 4-3. Sensitivity (left) and reaction time (right) for low and high load streams in the RSVP cross task. Both measures are significantly different between load conditions. Error bars represent \pm SEM.	109
Figure 4-4. Orientation offset direction discrimination accuracies. Error bars indicate \pm SEM across participants.	110
Figure 4-5. Mean GLM parameter values for visually responsive voxels in V1 (left), V2 (middle), and V3 (right) under low and high perceptual load conditions.	111
Figure 4-6. Population-wide VTFs (across 14 participants) in each load condition calculated using V1 voxels . VTFs are fitted with Von Mises functions. Error bars indicate \pm SEM across participants.	113
Figure 4-7. Gaussian kernel density estimates of the V1 VTF amplitude difference (left) and bandwidth difference (right) null distributions.	114
Figure 4-8. Population-wide VTFs (across 14 participants) in each load condition calculated using V2 (left) and V3 (right) voxels.	115
Figure 4-9. Distribution of orientation preferences in V1 (left), V2 (middle), and V3 (right).	115
Figure 4-10. MVPA classification results for patterns extracted from V1 (top), V2 (bottom-left), and V3 (bottom-right) activity.	116
Figure 5-1. Example frames from captured video in Brussels city centre.	125
Figure 5-2. Google maps screenshots of the 2 planned routes in central Brussels.	126
Figure 5-3. On the left, the types of road situation in the dataset by frequency, and on the right the number of videos per location group size (e.g. there were 488 videos matched with one other video at the same location).	128
Figure 5-4. Layout of experimental interface.	130
Figure 5-5. Histogram of perceptual load values as estimated by the TrueSkill algorithm.	132
Figure 5-6. Correlation of the perceptual load values across the whole dataset between comparison rounds.	133
Figure 5-7. Example dense trajectories in our dataset. The red points indicate sampled trajectory positions in the current frame, and the green trails indicate their locations in previous frames.	134
Figure 5-8. The lognormal distribution, parameterised by a mean, μ , and standard deviation, σ	138
Figure 5-9. Progress of the SMBO algorithm in terms of model variance explained across the 500 algorithm iterations. Blue dot markers represent a new best configuration being discovered by the algorithm.	139
Figure 5-10. Each blue marker represents a test set exemplar - it's position on the x-axis is the ground-truth TrueSkill estimate of perceptual load, while the y-axis position is its predicted perceptual load.	140
Figure 5-11. Progress of SMBO for SVR with nonlinear χ^2 kernel fusion of IDT channels.	141
Figure 5-12. SVR performance of original IDT pipeline applied to the prediction of perceptual load. Each blue dot represents a test set exemplar.	142
Figure 5-13. Progress of the SMBO algorithm in terms of model variance explained across the 500 algorithm iterations.	143
Figure 5-14. Ridge regression performance on held out test set for IDT+C3D descriptor combined using the linear kernel, where each blue marker represents a test set.	144
Figure 5-15. Progress of SVR with linear kernel channel fusion during 500 iterations of SMBO.	145
Figure 5-16. SVR with linear kernel, performance on the validation set after learning on full training set. Each blue marker represents a validation set exemplar.	145

Figure 5-17. Progress of SMBO for ridge regression with nonlinear kernels on IDT and C3D features	147
Figure 5-18. Predicted vs. actual load value plot for a ridge regression model using nonlinear multichannel kernel	148
Figure 5-19. SMBO progression for SVR with nonlinear channel kernels	149
Figure 5-20. Predicted perceptual load values of validation set examples plotted against actual values for the nonlinear multichannel kernel SVR model	151
Figure 5-21. SMBO progression over 500 iterations for optimising regularisation and channel weights for multichannel kernel computation and ridge regression	152
Figure 5-22. Best performing configuration of ridge regression with individual channel weights on the held-out validation set	153

Tables

Table 5-1. Features of the driving scene used to describe captured video and partition into individual sequences.....	127
Table 5-2. Tunable hyperparameters of the pipeline configurations in the current experiments.	137
Table 5-3. Best configuration of IDT hyperparameters found after 500 SMBO iterations	141
Table 5-4. Best found hyperparameters after 500 SMBO iterations for a ridge regression model with nonlinear multichannel kernel.....	148
Table 5-5. Best hyperparameter set found with TPE-based SMBO for multichannel nonlinear kernel with SVR regression.....	150
Table 5-6. Hyperparameters optimised after 500 iterations of SMBO for nonlinear kernel ridge regression with variable channel weights.	152
Table 5-7. R2 scores of the best performing models in each kernel-regression configuration	154

Acknowledgements

Firstly I would like to thank my supervisor Nilli Lavie for being a great motivator and guide during our time working together. Without your encouragement this thesis would not have been possible. Also a special thanks to Sam Schwarzkopf, for generously devoting time to explaining neuroimaging methods and designing experiments, whose enthusiasm for the subject is infectious. I'm also indebted to Gabe Brostow, for many helpful discussions and methodological ideas on the computer vision side of my research. I'd also like to thank Jonas Ambeck-Madsen of Toyota Motor Europe for believing in the project and being a great industrial partner and manager during my time spent interning at TME (also thanks on that front to Patrick Sauer from TME). Of course, a big thank you to all the residents of the attention lab, past and present: Moritz, Jake, Kate, Josh, Ana, Freya, Dana, David, Mahmoud, Rashmi, Alina, Theresa, Tim, Mike, Fintan, and Andrea. To my family and friends, who may be as relieved as I am that the thesis is written, a big thanks for the encouragement. I would especially like to thank Charlotte for standing by me during this stressful time and making me happy.

This research has been supported by a Toyota Motor Europe and University College London IMPACT award.

Preface

Seemingly obvious and salient objects can go completely unnoticed when a person's attention is directed towards a demanding task. This phenomenon of inattention blindness can be seen in the famous 'Invisible Gorilla' demonstration (Simons & Chabris, 1999), where an observer's task is to watch a video of a group of people passing a basketball amongst themselves and count the total number of successful passes. After completing the task, it is then revealed to the observer that in fact a full-sized adult in a gorilla suit had walked amongst the group, conspicuously pounding their chest while facing the camera, before exiting stage right. In the original study, slightly over half of the observer's did not notice the gorilla when viewing the video. There are obvious implications for safety in such cases of visual failure, however. For example, when driving a car or piloting a plane, failing to notice a crossing pedestrian, another road user, or important sign or signal could have potentially serious consequences. Indeed, a Department for Transport report (Brown, 2005) found 'failing to see' to be the third most commonly reported contributory factor to road accidents in Britain.

A major determinant of inattention blindness is the perceptual load of the task being completed (Cartwright-Finch & Lavie, 2007); when a task requires high levels of perceptual processing (e.g. to identify a target amongst many visually similar objects), reports of awareness of other stimuli are reduced significantly. High perceptual load has also been found to reduce detection sensitivity for random objects (Macdonald & Lavie, 2008), flickering lights (Carmel et al., 2007), and even across modalities, such that high visual demands reduce detection of auditory tones (Raveh & Lavie, 2014). These effects have been

explained in the load theory of attention through exhausting capacity; simply leaving no perceptual resources to perceive other stimuli¹

The underlying neural mechanisms by which increased perceptual load induces these deficits remains unclear however, with evidence mostly limited to behavioural and psychophysical measures, alongside neuroimaging findings of general attenuation of visuo-cortical activity under high perceptual load (e.g. Rees, Frith, & Lavie, 1997; Schwartz et al., 2005). However, the reported perceptual deficits may also be explained, at least in part, by degraded neural representation or selectivity for fundamental visual features such as colour, motion direction, and orientation at the earliest stages of visual processing; a possibility suggested by analysis of behavioural responses to orientation gratings (Stolte et al., 2014), but as yet not identified in visuo-cortical neural activity.

There is also uncertainty regarding the antecedents of perceptual load; that is, what elements or features of a task dictate the amount of perceptual processing required to complete it? And is it possible to predict from the visual information present in a task how much processing is required? Currently, perceptual load has traditionally been defined operationally by example (e.g. Lavie & Tsal, 1994; Lavie, 1995; Lavie, 2005), for example feature-conjunction search being higher load than single feature search, or target search amongst distractors being higher load when the number of distractors is increased. Recent modelling work by Roper et al. (2013) has expanded upon this somewhat, showing that several visual features of a task can be predictive of perceptual deficits for secondary stimuli, although this work was constrained to austere

¹ The extant limit on perceptual capacity is what requires attention (see Lavie, 1995; Lavie & Tsal, 1994). In fact, in tasks of perception, the concepts of a limited capacity attention or limited capacity perception are synonymous, as are the terms perceptual load and attentional load. I therefore use these interchangeably throughout the thesis

laboratory stimuli, where visual characteristics of the task were hand-labelled by the experimenter, (e.g. the letters C and T are not similar, whereas the letters L and T are of 'medium' similarity). It is clear that such an approach breaks down when faced with real-world perception, which operates in visually complex and dynamic conditions, where classifications of stimuli into simple object categories or similarity groups are not readily available. For perceptual load to have functional applications in the real world, such as identifying situations where load-induced inattention blindness is likely to occur, strides must be taken to estimate a priori, from task definitions and current visual information, the perceptual demand required in highly complex visual tasks.

The work presented in this thesis therefore aims to describe neural mechanisms by which increased perceptual demands of a given task degrade perception for other stimuli, and whether perceptual load itself can be estimated in a complex real-world task by analysing information in the visual field. In Chapters 3 and 4 the effect of perceptual load on the encoding of orientation in early visual cortex is investigated with modern functional magnetic resonance imaging (fMRI) methods such as multivariate pattern analysis (MVPA) and voxel-based tuning function (VTF) analysis. These methods allow the measurement of orientation-specific representational content at the level of visual areas (i.e. distributed patterns of activity) and at the neural population level (i.e. within voxels); representational content can be calculated under conditions of high and low perceptual load to identify neural loci of load-induced perceptual degradation. In Chapter 5 a modelling approach rooted in computer vision and machine learning is employed to produce a direct mapping between the raw visual field and perceptual load during the task of urban driving. This work extends research on perceptual load theory to practical applications in a real-world task, while the novel method introduced is applicable with slight modifications to a number of domains where estimating the likelihood of distraction or inattention

blindness can be critical. The following general introduction reviews the relevant empirical findings which provide the starting point for the experiments and analyses undertaken in this thesis.

1 General introduction

1.1 Visual attention and perceptual load theory

The total quantity of information available to our perceptual systems cannot be perceived due to the finite processing capacity of those systems; indeed, a calculation by Lennie (2003) implies that only 1% of the brain's neurons can be significantly active at any given time. Therefore, there must be a process by which a useful subset of all available information is selected, namely *attention*. While models of attention have become increasingly more sophisticated in the last 50 years, transforming from linear pipeline models (e.g. Broadbent, 1958) to those containing more complex structures such as feedback loops and parallel processing (e.g. Itti, Koch, & Niebur, 1998), a central facet has been debated throughout: at which stage during perceptual processing does attentional selection occur? Broadly two camps emerged, those espousing *early selection* and those espousing *late selection*. Under the early selection view, selective attention acts to exclude irrelevant information at the earliest stages of visual processing on the basis of rudimentary visual features of the stimulus (e.g. motion, colour, and orientation), before full perception and meaning can be extracted from the stimulus (e.g. Broadbent, 1958; Sperling, 1960). In contrast, late selection theories posit that all information is processed to full perceptual level (e.g. up to full recognition and semantic understanding of objects) and irrelevant information is discarded from post-perceptual processes such as response selection and memory (e.g. Deutsch & Deutsch, 1963; Allport, 1993; Eriksen & Eriksen, 1974). The confusion over conflicting experimental results and theories led Allport (1993) to suggest that the debate may never be resolved.

Lavie and Tsal (1994) proposed a resolution to this debate with the theory of perceptual load. This posits a role for the relevant task's *perceptual load* in the extent in which task-irrelevant sensory information is processed. A central tenet of the theory is the automatic and mandatory perception of information until a perceptual capacity is exhausted. The theory thus incorporates both early and late selection views: when the perceptual load of the relevant task is low (e.g. visual search for a target amongst a few easily distinguishable distractors), task irrelevant information is perceived; the task-relevant information being selected at a later stage. However when the load of the primary task is high and exhausts the perceptual capacity itself (e.g. visual search for a target among many similar distractors), the processing of irrelevant information is instead reduced at the earliest stages of perception. In order to operationalise the concept of perceptual load, Lavie (1995) put forward a definition in which an increase in the perceptual load of a task is precipitated by either 1) an increase in the number of task-relevant items it is necessary to perceive in order to perform the task or 2) an increase in the amount of perceptual processing required to perceive the task-relevant stimuli while viewing the same display (e.g. a target defined by a conjunction of basic features compared to a single feature).

This definition also leads to an important distinction within load theory between perceptual load and general task difficulty. Lavie and De Fockert (2003) conducted experiments in which target stimuli needed to be identified amongst non-targets in a canonical response-competition design (e.g. Eriksen & Eriksen, 1974; Lavie, 1995, 2005). Three conditions were presented in each experiment: 1) a high perceptual load condition where targets were identified amongst several non-targets, 2) a low-load condition where targets were presented alone (along with the competing irrelevant distractor), and 3) a low-load *degraded* condition where target stimuli were physically degraded in some fashion, for

example by reduction in size and contrast or by reduced display duration (see Figure 1-1).



Figure 1-1. Example displays used in each of three conditions by Lavie and DeFockert (2003). Each display here contains an incompatible distractor. The left panel shows a high load display, the middle panel a low-load display, and the right-most panel a low-load display with a degraded target of reduced contrast (adapted from Lavie and DeFockert, 2003).

They found that while high perceptual load and sensory degradation of the target increased task-difficulty, as measured by target identification accuracy and response latency, the physically degraded low-load displays did not show a reduced response-competition effect due to irrelevant distractors, as found for displays high in perceptual load. The results indicate that such manipulations of perceptual load influence distractor processing independently of sensory 'data limits' (Norman & Bobrow, 1975), and in fact reflect a bottleneck in attentional capacity.

Many conflicting findings which contributed to the early vs. late selection debate can be reconciled using this operational definition of perceptual load, within the terms of perceptual load theory. For example, experiments using variations of the Stroop task (e.g. Eriksen & Eriksen, 1974; Gatti & Egeth, 1978) found that irrelevant flankers were often identified correctly, providing evidence for a late selection approach. However, these experiments used displays with low levels of perceptual load, usually no more than a single target and a single distractor. Equivalently, much of the earlier research which led to an early selection interpretation (e.g. Snyder, 1972; Treisman & Riley, 1969) used tasks of high

perceptual load, with high numbers of distractors or high similarity between targets and distractors.

Before I turn to review studies into the effects of increased perceptual load on visual perception, it is important to couch the concept of perceptual load and its effects in the combined theory of perceptual and *cognitive control* load (for reviews see Lavie, 2005; 2010). In this view, there are two complementary mechanisms that enable selective attention; the first is the mandatory spill-over of attention to irrelevant stimuli when the perceptual load level of the primary task is low – this can be seen as an early selection mechanism. The other, late-selection mechanism posits that cognitive control processes dictate the extent to which irrelevant information can be eliminated from further processing. This form of selection is dependent upon higher-level executive functions, such as working memory (WM), to maintain task-dependent stimulus priorities and ensure that low-priority information does not enter awareness and guide behaviour. Therefore, the full model predicts that when the task at hand is highly demanding of cognitive control processes, thus inhibiting their ability to monitor and maintain stimulus priorities, irrelevant and unattended information is likely to undergo further processing; a loading effect opposite to that of perceptual load.

To investigate this, De Fockert et al. (2001) varied cognitive control load with a digit memorisation task which recruited WM: under low WM load participants were required to memorise a set of digits in a numerically ascending order, while in the high WM load condition, the digits were memorised in a random order. Concurrently, a name categorisation task was completed, where celebrity names had to be categorised while congruent or incongruent facial images of the same celebrities were presented in the periphery. It was found that face distractor effects on reaction times were greater when cognitive control load

was higher, and that neural activity in the fusiform face area (FFA; an area associated with the perception of faces, see Kanwisher, McDermott & Chun, 1997) was *increased* under high WM load. These effects are indeed the opposite of those seen for similar increases in perceptual load, highlighting an important distinction between the effects of perceptual and cognitive control process demand on perception.

1.1.1 Behavioural evidence for effects of perceptual load

There have been many studies investigating the effects of increased levels of perceptual load on the processing of distractor stimuli. A canonical manipulation of perceptual load uses the response competition paradigm (Eriksen & Eriksen, 1974). Lavie and Cox (1997) found that in conditions of low perceptual load, reaction times to the task are increased when an irrelevant distractor is incongruent to the target; an effect that disappears when the perceptual load of the task is high (see Figure 1-2). The result implies that under high perceptual load the processing of the irrelevant distractor is reduced, as the identity of the distractor does not interfere with the relevant task response.

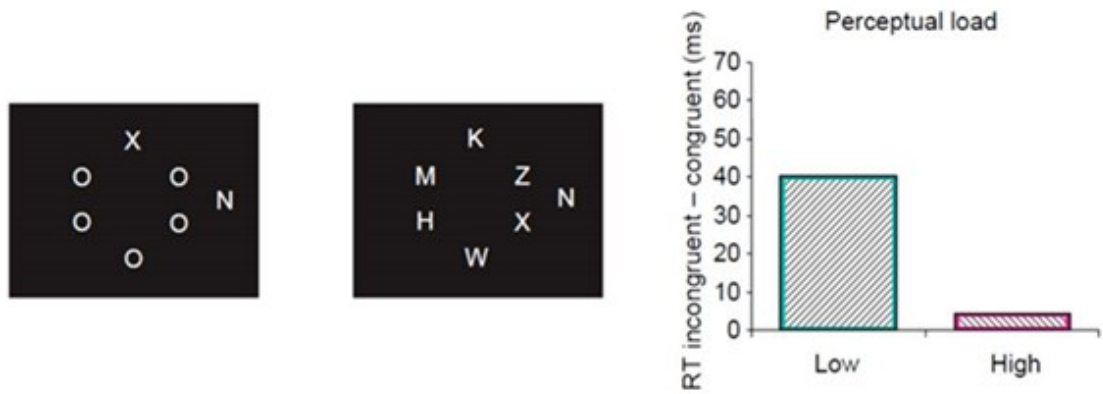


Figure 1-2. Perceptual load manipulation in a response competition design (e.g. Lavie & Cox, 1997). Displays of either low (left) or high (middle) perceptual load are presented, combined with an irrelevant distractor letter (here N, the rightmost letter in the displays). The subject makes a speeded response to one of two predetermined targets (here X or N). Under low load, when the irrelevant distractor is incongruent with the actual target and visual search can proceed efficiently (e.g. target X, distractor N) reaction to the target takes longer than when the distractor is congruent, where target search is inefficient (e.g. target X and distractor X). However, this effect is eliminated under high load

Forster and Lavie (2008) extended these findings to the processing of completely irrelevant distractors, in which the distractor stimuli bore no similarity to the perceptually relevant items of the primary task in terms of visual appearance, meaning or location. (see Figure 1-3)

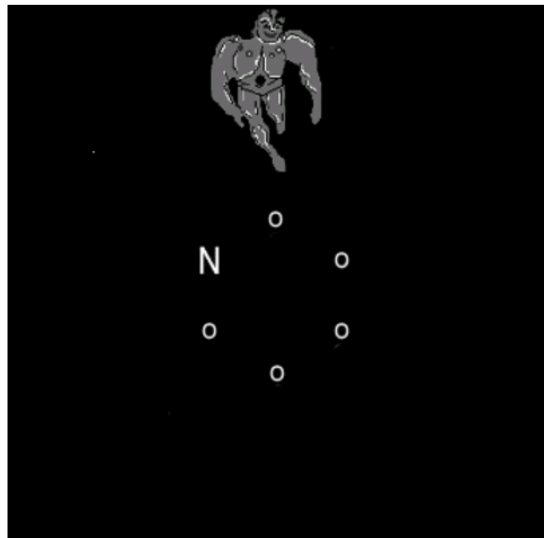


Figure 1-3 An illustrative example of a low-load display used by Forster and Lavie (2008) containing a completely irrelevant distractor; a cartoon character

An identical pattern of results was found to that of Lavie and Cox (1997) – a distractor effect on primary-task reaction times under low load was eliminated under conditions of high perceptual load. The finding broadens the scope of the effects of perceptual load to a form of distractibility common in everyday life, where distractors are unlikely to be physically or meaningfully related to the task-at-hand. The dependent variable used in these experiments - response reaction time - is adequate to confirm differences in the gross extent of distractor processing under varying levels of perceptual load; however the nature of this difference at the perceptual level is not attainable with such designs. To elucidate perceptual effects in more detail, that is, *how* load changes our perception, Macdonald and Lavie (2008) investigated the effect of perceptual load on stimulus detection sensitivity in a dual-task design. The participants' primary task was a letter-search task presented at fixation which could be either low or high load. There was also a simultaneous detection task in the periphery, where participants were instructed to respond when a certain critical stimulus was presented. They found that sensitivity (as measured by d') to the presence of the critical stimulus was significantly reduced under high load, demonstrating that perceptual load acts to reduce our perceptual ability to detect task-relevant information, beyond any potential simple effects of response bias induced by load.

Carmel, Saker, Rees, and Lavie (2007) investigated the role of perceptual load in the perception and detection of a rapidly flashing stimulus. They presented participants with either a low or high load letter search in the periphery, while a red flickering stimulus was presented at fixation. The stimulus was presented on or very near to each participant's critical flicker fusion (CFF) threshold – the temporal frequency of a flickering light which is equally likely to be perceived as flickering or fused (i.e. a continuous 'lit' percept). They found that physically identical flickering stimuli were less likely to be perceived as flickering when the

letter search was high load. Furthermore, when the task was adapted to a two-interval forced choice paradigm, psychophysical measures of flicker detection sensitivity were shown to be significantly lower under high perceptual load; establishing that perceptual load can affect the perception of stimulus features in the temporal domain. Raveh and Lavie (2014) have recently extended this work to inter-modality effects of perceptual load, finding that high load in a visual task significantly reduces detection sensitivity for auditory tones, providing support for a shared, finite, attentional resource across modalities. Stimulus detection studies have also extended the scope of load research beyond effects of primary task load on *irrelevant* visual stimuli; showing that perceptual load modulations in a primary task can alter the perception of task-relevant stimuli presented in a dual-task paradigm; if attentional demands of the secondary task are kept constant, then performance changes in this task can be attributable to load manipulations in the primary task.

In Lavie, Lin, Zokaei, & Thoma (2009) subjects performed a letter-search task while a wide range of meaningful, realistic, but task-irrelevant distractors (e.g., a picture of a car), were presented in the periphery. After completion of the attention task, a surprise memory-recognition task was presented. Results showed that even when distractor stimuli were presented directly at the position of subject's fixation, they could only recognize having previously seen those objects when the attention task at initial viewing involved was low load, and in such conditions, recognition was possible even under different views of the object. However, recognition memory fell to *chance* levels when the initial viewing was under high perceptual load. The results imply that when perceptual resources are exhausted by the processing demands of a certain task, even view-dependent object representation are unable to be formed and stored; in contrast to low load situations, where a mandatory spill-over of perceptual resources leads to the creation of rich and persistent view-invariant

representations of task-irrelevant objects. It is worth noting here however, that the objects used in this experiment were presented in isolation, free of real-world context; to date there has been no studies utilising real-world complex scenes as stimuli to induce, or measure the effects of, perceptual load modulations.

1.1.2 Neuroimaging evidence for effects of perceptual load

With the advent of functional magnetic resonance imaging (fMRI) techniques, many studies have attempted to uncover the neural correlates of perceptual load effects. These methods allow the indirect measurement of brain activity, as given by the blood oxygen level dependant (BOLD) signal, attributable to irrelevant stimulus processing. Experiments can therefore be designed similarly to earlier behavioural studies, with a primary load task and a distractor presented, and differences in brain activity due to visual stimulation a being attributable to the manipulation of perceptual load.

Yi et al. (2004), using functional magnetic-resonance imaging (fMRI), investigated the effects of perceptual load on neural activity elicited by complex, real-world images of houses and natural scenery. At fixation participants completed an *n*-back face-recognition task which was varied in its perceptual and cognitive control demands; through introducing noise into the face images and by increasing *n* in the *n*-back task, respectively. They localised activity in the parahippocampal place area (PPA; Epstein & Kanwisher, 1998), an area in medial temporal cortex which responds selectively to imagery of places and scenes. The results showed that activation in the PPA due to the imagery of houses and scenery was significantly reduced when the foveal task was high load, providing evidence that increased perceptual load in a visual task directly

attenuates levels of neural activity for stimuli outside the focus of this task. Moreover, this effect was not present for increases in working memory load, supporting claims of complementary roles for perceptual and cognitive control demands in perceptual tasks. In another fMRI experiment, Pinsk, Dongier, and Kastner (2004) found similar load effects on the neural processing of complex imagery, however they found modulation of activity in the inferior temporal area TEO as well as visuo-cortical area V4; increases in the perceptual load of a task were shown to affect the processing of stimuli in *the visual cortex itself*.

In Pinsk, Dongier, and Kastner's (2004) design, differences were not found between load conditions for stimulus-induced neural activity in lower-level visual areas such as V1 – this may be due to the complex nature of their stimuli however, other work using highly salient imagery likely to stimulate earlier areas (e.g. flickering checkerboards, high-contrast motion patterns) has reported effects of perceptual load on neural activity. For example, O'Connor et al. (2002) found modulations of activity in the lateral geniculate nucleus (LGN), the main connection from the optical nerve into the early visual areas of the occipital lobe, while in an EEG study, Parks et al. (2011) found a reduction in primary visual cortex (V1) signals for peripherally presented flickering checkerboards when perceptual load in a foveal task was increased. In an fMRI design, Rees, Frith, and Lavie (1997) investigated the activity due to irrelevant distractor motion in early motion-selective visual areas V1, V2, and V5/MT. They manipulated the load level of a central task whilst presenting a task-irrelevant pattern of high-contrast moving dots in the periphery. Across two conditions, the dot-pattern could either be static or in constant motion. They found that neural activity attributable to the motion patterns was significantly attenuated under high load.

Schwartz et al. (2005) reported perceptual loading effects on activity elicited by a simple but salient flickering checkerboard stimulus, across multiple early visual areas. They presented participants with a central rapid successive visual presentation (RSVP) task consisting of a stream of crosses, which varied in two dimensions: the colour of the cross and whether it was presented upright or inverted. In the low load version of the task, the subject had to respond to any *red* cross irrespective of its inversion state, while in the high load version participants responded to either *upright yellow* or *inverted green* crosses when presented. Simultaneously participants were presented with flickering checkerboards peripherally, designed to stimulate visual cortex.

They found that areas traditionally associated with attentional demands, such as the frontal gyrus and parietal lobule, were more active under high load as compared to low. They also found multiple clusters of voxels located in visual cortex which were significantly less active under high load, indicating the reduced processing of irrelevant peripheral distractors. After retinotopically delineating early visual areas (V1 to V4), it was confirmed that activity under high load was significantly lower across these areas, with the load effect increasing in later areas (V2 to V4). These results suggest that perceptual load acts to suppress the neural response to visual stimulation at the earliest stages of visual cortex. Bahrami, Lavie, and Rees (2007) extended these findings to distractor stimuli that were rendered perceptually invisible through continuous flash suppression (Tsuchiya & Koch, 2005); finding that neural activity induced even by stimuli of which we are unaware is attenuated by increasing perceptual load of a primary task.

Therefore, much of the earlier behavioural research has been corroborated and, in-part, explained by neuroimaging research: neural activity throughout the visuo-cortical hierarchy due to distractor stimuli is reduced under conditions of

high perceptual load. However the explanation has been limited to modulation of gross activity levels under different load conditions, without expanding upon the exact nature of that modulation.

1.2 Modulation of feature-specific representation

Much behavioural and imaging evidence therefore supports load theory's basic tenet, that stimulation irrelevant to the primary task undergoes reduced processing when the task heavily loads the perceptual system. A natural next step is to investigate the mechanism of these differences, much like the shift in behavioural experiments to uncover deeper perceptual effects of load through applying psychophysical designs. To use the example of Schwartz et al.'s (2005) experiment described above, although it is found that the gross activity of visual cortex induced by a peripheral checkerboard stimulus is modulated by load, the experiment offers no insight as to how representations of the basic features comprising the checkerboard are affected by load. Might the response along dimensions of its low-level features, such as colour or orientation, be modulated by load also?

Recently developed fMRI analysis techniques allow the measurement of such phenomena physiologically. Multi-voxel pattern analysis (MVPA; see Chapter 2 for detailed description) is able to infer (or *decode*) the orientation of gratings shown to a subject based on distributed activity across a visual region (Kamitani & Tong, 2005; Haynes & Rees, 2006), and presents an avenue for measuring the amount of feature-specific information contained within distributed cortical representations. For example, to investigate the effect of perceptual expectation on orientation representation, Kok et al. (2012) measured the representational content of distributed activity in V1 to oriented grating stimuli during a change

detection task. The change detection task consisted of two oriented gratings being presented sequentially to the subject, the second being slightly rotated relative to the first, in either the clockwise (CW) or counterclockwise (CCW) direction; the subject was required to respond as to which direction the second grating had been rotated. Preceding the presentation of the first grating was an audio signal which on 75% of the trials correctly indicated the orientation of the first grating (either 45° or 135°) whilst on the remainder of the trials it indicated the incorrect, other orientation; in this way the perceptual expectations of the subject were manipulated. It was found that on trials with a correct aural cue the orientation of the subsequent grating orientation could be decoded from V1 activity with significantly higher accuracy than in trials with a deceptive aural cue, showing that certain behavioural trends associated with an experimental manipulation (perceptual expectancy in this case) can be traced using MVPA methods to modulations of feature-specific representations at the earliest stages of visual processing.

With regard to the effect of attention on neural representational content, Kamitani and Tong (2005) manipulated the application of feature-based attention by presented participants with superimposed oriented gratings of two orientations (45° and 135°). Participants fixated centrally but were cued to attend to one or the other orientations; using MVPA they found significantly higher decoding accuracy in area V1 for the attended orientation compared to the unattended. Using an identical orientation-superposition design, Serences et al. (2009) analysed the imaging data by constructing voxel-based tuning functions (VTFs; see Chapter 2 for methodological details) from BOLD responses to the stimuli, population-scale analogues of neuronal tuning curves. This analysis method pools the tuning properties of small neural populations (as measured by voxel BOLD response) across an area rather than extracting the distributed representation as with MVPA, and allows the calculation of fine-scale

neural tuning parameters such as response amplitude, response bandwidth, and feature preference. When attention was cued to one of the orientations in the superposition it was found that the preferred orientation (i.e. that which elicited maximal activity) of neural populations in early visual areas was shifted towards the attended orientation. Therefore there now exist several techniques within the fMRI domain to investigate the effect of attentional manipulations on feature-specific neural representations.

It is established that high perceptual load is a major determinant of reduced perception of stimuli outside the focus of an attended task; however the precise mechanisms of this suppression remain unclear. While behavioural studies in the domain of orientation perception (Stolte, Bahrami, and Lavie; 2014) indicate the nature of low-level feature-selective modulations by perceptual load, these effects have not been corroborated by measurement of neural responses themselves; it is therefore unknown whether the reported psychophysical modulations can be traced to the earliest stages of visual processing, and further, whether modulations of visuo-cortical activity follow the same trends as response modulations at the behavioural level.

Recent work by Stolte, Bahrami, and Lavie (2014) investigated this psychophysically using a noise-masking dual-task paradigm. While simultaneously completing a primary visual search task, which could either be high or low load, subjects were presented with displays containing a vertically oriented Gabor patch embedded in a circular noise mask; the patch was at horizontal centre, but could be shifted either slightly above or below vertical centre, constituting the two states of a two alternative forced choice task. The participant was instructed to first respond to the load task, and then immediately respond to the orientation detection task, indicating whether the Gabor patch had been displayed above or below vertical centre. Within each load condition,

two variables were manipulated to produce orientation tuning curves: the contrast and orientation of the noise mask. A set of seven orientations were used (0° , 8° , 16° , 24° , 32° , 64° , 90°), whilst an adaptive staircase method (QUEST, Watson & Pelli, 1975) was used to obtain a 75% accuracy threshold for each orientation. Curves were then constructed with the dependant variable being contrast value at 75% detection accuracy (see Figure 1-4). Under high load, tuning curves showed increased contrast threshold, indicating reduced gross neural activity, as well as increased bandwidth. This result therefore behaviourally demonstrates a role for perceptual load as a determinant of orientation selectivity for secondary stimuli.

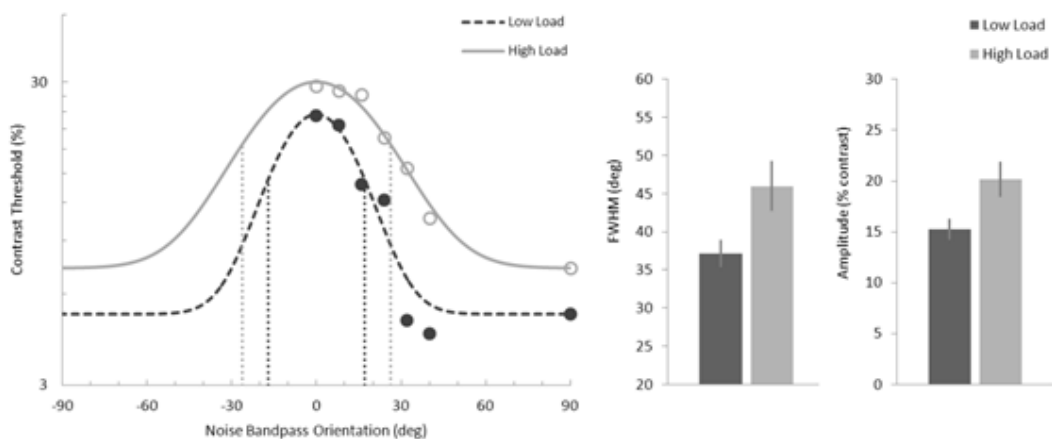


Figure 1-4. Psychophysical tuning modulation due to perceptual load. On the left, characteristic orientation population tuning curves from a participant under both load conditions. Notice increased contrast threshold and broadened tuning under high load. On the right, averages across all participants are given, showing the overall decrease in bandwidth and increase in amplitude due to the load manipulation (from Stolte, Bahrami, & Lavie, 2014).

Although no research has previously looked at the impact of perceptual load on orientation tuning specifically, one recent line of work represents a step towards understanding the impact of perceptual load on low-level feature encoding. De Haas et al. (2015) investigated the effect of perceptual load on population receptive fields (pRFs; Dumoulin & Wandell, 2008) in an fMRI experiment. pRFs are a population-level analogue of neuronal receptive fields and characterise the response properties of voxels to stimulation across the visual field. A single

voxel's pRF is represented by a 2-dimensional Gaussian function which is centred at the location in the visual field where visual stimulation elicits the maximal BOLD response. De Haas and colleagues extracted pRFs for voxels in early visual cortex under conditions of low and high perceptual load, and subsequently compared parameters of the fitted Gaussian pRF models. They found that under high load, voxels with pRFs centred parafoveally were responsive to more of the visual field (i.e. the Gaussian functions had an increased spread) relative to low load. The finding indicates that task-related perceptual load can influence the processing stimulus location, however no work up to now has investigated primitive feature encoding, such as orientation or motion, under load. Therefore, in Chapter's 3 and 4 we investigate the effect of perceptual load on the orientation-specific representational content of neural populations in early visual cortex.

1.3 The causes of perceptual load

In the original formulation of perceptual load theory, Lavie (1995) put forward an operational definition of perceptual load to resolve the conflict between early and late selection theories of visual attention. By this operational definition, an increase in the perceptual load of a task is precipitated by either 1) an increase in the number of task-relevant items it is necessary to perceive in order to perform the task or 2) an increase in the amount of perceptual processing required to perceive the task-relevant stimuli while viewing the same display (e.g. a target defined by a conjunction of basic features compared to a single feature; Lavie, 1995; Schwartz et al., 2005).

Experimental manipulations based in the operational definition of load set forth by Lavie in 1995 have been shown to modulate attentional systems, rather than

general effects of difficulty (e.g. Lavie & De Fockert, 2003). However, while these manipulations have been employed numerous times in order to investigate effects of increased perceptual load on the perception or processing of, there currently exist few attempts to produce an objective, continuous definition of perceptual load *per se* in terms of specific task and stimulus characteristics. Furthermore, these attempts concern only the canonical response-competition demonstration of perceptual load effects.

A recent study aimed to produce *a priori* characterisation of task-induced perceptual load through recourse to the theory of biased competition (Desimone & Duncan, 1995). At the base of biased competition theory lays the assertion that objects compete for representation and processing at higher levels of visual processing, being as it is impossible for each visual object to be represented simultaneously due to the extended nature of visuocortical receptive fields. Which object is represented by a certain cell or group of cells is then dictated by modulatory top-down influences, such as selective attention, which *bias* the competition between visual objects within a receptive field towards object attributes relevant to the current task or situation. Torralbo and Beck (2008) applied biased competition theory to investigate a potential explanatory factor for perceptual load effects found in countless response-competition experiments, and provide a potential mechanism for assessing the perceptual load of a task given its definition and stimulus characteristics. They implemented the canonical response-competition paradigm; however they varied the 'density' of the letters in the letter-search rather than perceptual load explicitly (see Figure 1-5).

	Compatible	Neutral	Incompatible
Low Density			
High Density			

Figure 1-5. Example displays from Experiment 1 by Torralbo and Beck (2008). Low-density displays (top row) contain a central distractor along with a letter-search task (here, targets are Xs), where the individual letters in the search task are separated by a constant distance. This is in contrast to the letters in the high-density condition, which are not separated. Note that all forms of display in this experiment would fall under the high-load condition in other perceptual load modulation experiments, given that there are several non-targets from which to distinguish the target

For both low and high density conditions a significant difference was found between letter-search reaction times on trials containing compatible and incompatible distractors, in keeping with the apparent high-perceptual load of all displays. However, this response-competition effect was modulated by the density of the displays, such that high density displays resulted in smaller distractor interference effects than low density displays. Torralbo and Beck (2008) explain this as a result of biased competition: high-density displays mean that stimuli compete for representation in early and intermediate stages of visual cortex (areas with receptive fields small enough to elicit such an effect), thus top-down selection must heavily bias perception to isolate the representation of the target letter, resulting in reduced perception of the distractor letter.

However, this mechanism and explanation stands in contrast to the earlier findings of Lavie and De Fockert (2003), who varied the eccentricity of the search task in a similar design, finding that as eccentricity increased – inducing

reduced visual acuity due to receptive field size increase – that greater distractor interference effects were observed; an opposite effect to that predicted by the mechanism of Torralbo and Beck. Lavie and De Fockert (2003) explained their findings in terms of stimulus salience: the degradation of the target acuity results in increased relative salience of the distractor, the distractor is therefore more likely to win the race to awareness. Therefore, at present it is unclear whether density in task displays provides a promising avenue for developing an independent measure of perceptual load.

Roper et al. (2013) conducted experiments into the effect of stimulus similarity on perceptual load effects in an effort to objectively quantify the perceptual load induced by response-competition displays. Their work follows on from Lavie and Cox (1997) who varied set size in the response-competition paradigm, finding that an increase in the number of visual search distractors leads to reduced irrelevant distractor interference - up to a certain threshold number of items, consistent with the exhaustion of finite attentional resources. They also investigated the effect of visual search difficulty independent of set size.

They found that congruency effects for the irrelevant distractor were greatest in the 'easy' search conditions, suggesting that the perceptual load of a response-competition task display may be inversely correlated with efficiency in the search task. Roper et al. (2013) investigated this relationship further, suggesting that previous tests of perceptual load have typically confounded target–distractor (T-D) similarity and distractor–distractor (D-D) similarity in the response-competition search displays. Low-load displays often employ targets that are perceptually distinct from homogeneous distractors, whereas high-load displays employ targets that resemble heterogeneous distractors; the two conditions thus differ across both T-D and D-D dimensions. In their study

however, Roper and colleagues prepared stimulus displays which varied along both dimensions independently (see Figure 1-6).

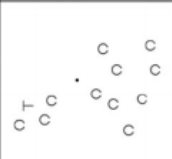
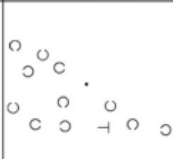
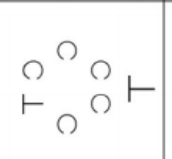
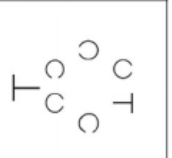


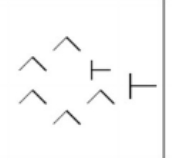
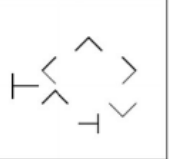
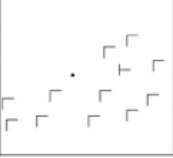
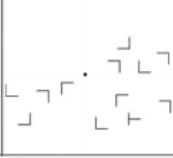
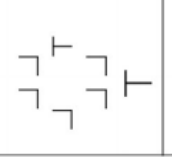
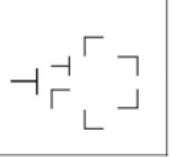
	Visual Search Task		Perceptual Load Task	
	High D-D	Low D-D	High D-D	Low D-D
Low T-D				
Medium T-D				
High T-D				

Figure 1-6. Displays used by Roper et al. (2013).

They found that the degree to which a certain T-D x D-D configuration produces efficient search, as measured by search slope in the search reaction time by set size relationship – a slope near 0 representing an efficient search stimulus set, was strongly correlated with the amount of irrelevant distractor interference induced by the same distractor set used in a response-competition perceptual load task. Roper and colleagues (2013) then constructed a multiple regression model using data across their experiments - incorporating several visual search factors such as search slope, search intercept, T-D similarity, and D-D similarity – to learn a mathematical, predictive model of irrelevant distractor interference in a perceptual load task. While search slope was the single best predictor of flanker effects, a model also incorporating search slope intercept was able to account for 98% of the variance in flanker interference effects (see Figure 1-7).

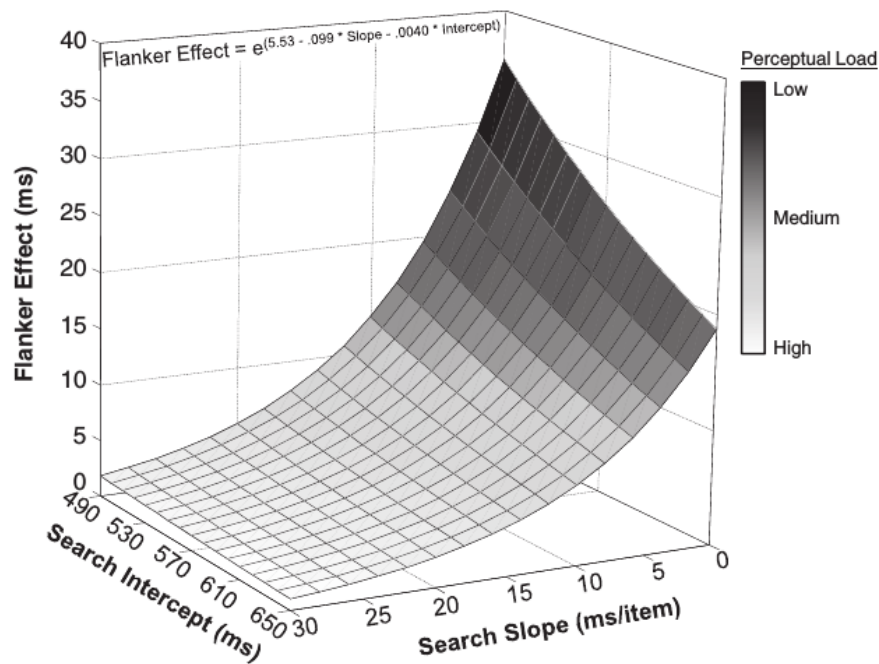


Figure 1-7. Predicted perceptual load induced response-competition task displays (as given by flanker interference effects). The model produces predictions using on an exponential relationship incorporating visual slope search and intercept values extracted from participants completing several visual search tasks independently

Roper et al.'s (2013) work therefore constitutes an objective, predictive model of perceptual load in the response-competition paradigm, based on performance in a related but independent visual search task. Importantly, performance in the visual search task is almost wholly determined by visual factors of the search display itself, namely T-D and D-D similarities; as such the model provides a method for quantifying the likely level of perceptual load induced by a task *a priori*.

While this type of work represents an important step in the development of perceptual load theory, the chosen task described by the model remains austere and relatively divorced from the function of attention and perception in the real-world: static, simple displays in comparison to the rich dynamic information processed in natural scenes and situations. Roper et al. (2013) do however introduce a promising data-driven methodology for exploring the antecedents of perceptual load for tasks in general, that of learning a

mathematical relationship between a task's visual features and the level of perceptual load induced by that task. In Chapter 5 of this thesis I therefore leverage and extend this methodology in an effort to predict perceptual load for a complex, dynamic, real-world task. The task chosen to model was urban driving, which involves the need to perceive a complex changing environment and attention is critical, while the definition of the task itself (which may be framed generally as collision avoidance) is familiar to many people. There is also a pragmatic and ethical motivation for this choice: a number of road accident surveys have shown that inattentive blindness is the third most frequently recorded contributory driver error (Department for Transport review of the 'looked but failed to see' accident causation factor; Brown, 2005). Indeed a study of naturalistic driving behavior revealed that inattention contributed to 78% of accidents (Klauer et al., 2005). The ability to identify driving situations which induce high perceptual load, and therefore estimate the related likelihood of blindness to safety-critical incidents, could establish an avenue for developing novel safety features such as driver intervention methods to maintain attention on the road.

1.4 Perceptual load in driving

Several studies have recently investigated the effects of increased perceptual load on driving performance measures in driving simulators (e.g. Redondo & Lee, 2009; Marciano & Yeshurun, 2012; 2015). For example, Marciano and Yeshurun (2015) manipulated perceptual load for subjects in control of a driving simulator, on the road itself by varying the density of vehicles in the immediate vicinity of the ego-car (i.e. the subject's car) and the density of pedestrians on the side of the roadway (see Figure 1-8).

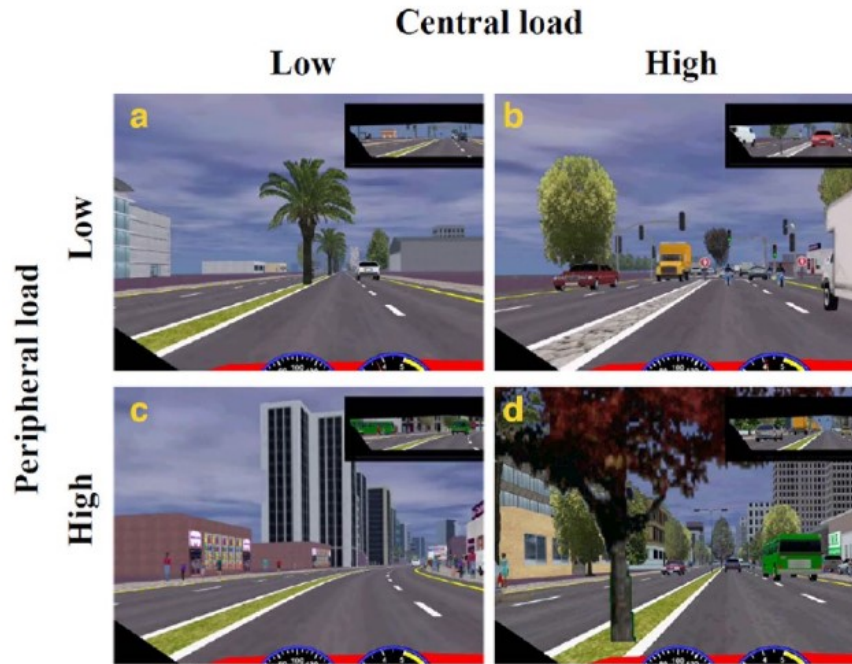


Figure 1-8. Screenshots of driving simulator in the experimental conditions employed by Marciano and Yeshurun (2015). Central load relates to the number of vehicles occupying the road surrounding the ego-car, while peripheral load was manipulated with the number of nearby pedestrians

Several measures of driving performance were collected during simulated driving under each of the four conditions, including average driving speed and reaction time to respond to pre-planned critical peripheral and central events such as a pedestrian suddenly crossing the road or the leading vehicle braking in front of the ego-car. Results showed that when central perceptual load was low, subjects' average speed increased, indicating an assumption that they could maintain safe driving at higher speeds. However, this was also accompanied by an unintuitive finding that drivers were less likely to respond to critical peripheral events in good time, seemingly due to the increased speed, perhaps highlighting a general shortcoming in the validity of simulator studies, where safe driving is not actually critical to the driver. The effect of central load on response to central critical events, and the effect of peripheral load on both types of critical event, was more normative however, with subjects being more likely to respond correctly under conditions of low perceptual load.

In another recent simulator study, Murphy and Greene (2016) investigated the effects of perceptual load on drivers' awareness for a pedestrian or large animal situated at the side of the road. Perceptual load was modulated by a task requiring the participant to indicate whether their vehicle would fit between two rows of vehicles lined up at the side of the road. In the low load condition, the space between the vehicles was obviously too large or too small for the ego-car to pass through, while in the high load condition the gap was only slightly too narrow or wide.



Figure 1-9. Example of low perceptual load in Murphy and Greene's (2016) simulator experiment. This represents a critical trial, as a pedestrian is placed at the side of the road.

On a random low and high load section of driving the pedestrian or large animal was presented to the side of the road; immediately after the participant had completed the critical trial (i.e. by driving through or around the two rows of parked vehicles) a prompt asked the participant whether they had noticed anything unusual in that section of driving. In line with previous laboratory manipulations of perceptual load, subjects' awareness for task-irrelevant stimuli was reduced when the perceptual load of the task-at-hand was high.

While these simulator studies mark an important research venture into the application of perceptual load theory to real world situations, the current state is

analogous to that in traditional laboratory research. Perceptual load has hitherto been modulated *a priori* by the experimenters through recourse to the original operational definitions of perceptual load, and the field is lacking a measure of the perceptual load of a driving situation without invoking the *effects* of perceptual load. Furthermore, while studies involving simulators claim a level of ecological validity, the visual quality of the simulation is often deprived – with highly planar, simplistic graphics (see above). In Chapter 5, I therefore attempt to address both of these shortcomings, producing a model which operates on real-world driving footage to estimate the perceptual load induced by the driving situation.

In adopting a data-driven model-fitting approach to this problem we are necessarily faced with a challenge with regard to obtaining data which captures visual features of driving along with an associated, assumed *ground-truth* measurement of perceptual load. In modelling a simple task such as the response-competition paradigm, for example, the properties of the task itself can be manipulated precisely and numerically (e.g. increasing the number of search distractors), and a direct objective dependent variable can be obtained (i.e. the reaction time interference effect of congruent vs. incongruent flanker stimuli), thus enabling the construction of a mapping between task features and the dependent variable (as in Roper et al., 2013). The greater inherent complexity and variability present in real-world driving however, precludes us from implementing such precisely defined conditions and manipulations. It is therefore apt to employ a semi-observational study design, in which natural driving footage is collected as experimental stimuli along general heuristics rather than within a tight factorial design, and associate with each driving situation a value of perceptual load in the driving task. Using this footage, labelled with perceptual load values, a regression model can be fit between fundamental spatio-temporal features of the dynamic scene and perceptual

load, resulting in a system able to estimate the level of perceptual load induced by a given driving scenario.

2 Methodological background

This chapter outlines and gives historical context to the methods used throughout the rest of the thesis. Chapters 3 and 4 employ functional magnetic resonance imaging (fMRI) techniques to investigate the effect of perceptual load on fundamental feature representation in visual cortex. A brief summary of basic fMRI technologies through to the modern data analysis approaches used in this work is therefore covered in the next section. Following this, the methods which underpin the modelling work of Chapter 5 are described, such as the extraction of semantics from natural imagery, nonlinear kernel regression, and the method of pairwise comparisons.

2.1 Measuring and interpreting brain activity

Introduced by Ogawa and colleagues (1990), functional magnetic resonance imaging (fMRI) is an extension of magnetic resonance imaging (MRI) to the measurement of blood oxygen levels in the brain. Its invention has revolutionised much of modern neuroscience, allowing the non-invasive measurement and localisation of neural activity *in vivo*.

2.1.1 Measuring activity

2.1.1.1 Magnetic resonance imaging

MRI utilises the phenomenon of nuclear magnetic resonance (NMR) to produce images of matter concentrations in the body. NMR is a physical phenomenon whereby atomic nuclei placed in a magnetic field absorb and re-emit electromagnetic (including radio-frequency; RF) radiation. The radiation which

can be absorbed must be at a specific, discrete, *resonant* frequency, which is dependent on the strength of the magnetic field applied to the nuclei and the magnetic properties of the nuclei. Once excited, the nuclei then re-emit the energy over time, in a process termed *relaxation*, which is detectable by sensors. The time taken for those excited atoms to return to their previous equilibrium state is the basis of MRI images: different materials and tissues affect the relaxation time of nuclei to a greater or lesser extent due to their differing intrinsic magnetic properties. These differences can therefore be measured and displayed in monochromatic images, where different materials may be clearly separated based on this measured relaxation time.

Since the human body is around 80% water or fat, the magnetic properties of hydrogen nuclei (i.e. single protons), which are preponderant in these materials, are often used for MRI. To create an MR *image*, it is necessary for these protons to be distinguishable by their spatial location. In the resting state, the set of hydrogen protons in a body will absorb the energy of an RF pulse. However, if a linear gradient magnetic field is applied across the body, then the protons will resonate at different frequencies, aligned with the difference in magnetic field strength. Then, only a single band of protons, localised in a *slice* in space, will respond to a given RF pulse. By varying the RF pulse frequency, information regarding the magnetic properties of tissues in slices across the body can be measured. The MRI signal can be isolated beyond the level of slices into into small cubes, termed *voxels* (or, volume elements). This is done by applying a gradient magnetic field during the measurement of re-emitted radiation from an excited slice, such that areas exposed to a lower field magnitude emit their radiation at a lower frequency. Discrete increases in the strength of this field partition the excited slice into a number of segments in 2D. When coupled with the information regarding slice position each voxel is therefore distinguishable as a small cubic volume in space: the relaxation

profiles of protons across each voxel can be measured and assembled into a 3D image of relaxation times. This image can then be used to delineate different types of tissue (e.g. isolate tumours within healthy tissue), or, in the case of fMRI, convey the amount of neural activity throughout the brain.

2.1.1.2 The BOLD signal

fMRI measures neural activity through the proxy of blood oxygenation. The blood-oxygen-level dependent signal (BOLD signal) was first measured by Ogawa et al. (1990) in mice, who found that brain structures became less visible during MRI acquisition when mice inhaled higher concentrations of oxygen. This intriguing effect was explained by the relative magnetic properties of oxygenated vs. deoxygenated blood. Deoxyhaemoglobin (i.e. deoxygenated red blood cells) is paramagnetic, meaning it is attracted to an external magnetic field, while oxyhaemoglobin (i.e. oxygenated red blood cells) is very weakly diamagnetic. This means that deoxyhaemoglobin induces inhomogeneities into the surrounding magnetic field which are not present in the vicinity of oxyhaemoglobin (Heeger & Ress, 2002). This magnetic contrast between deoxyhaemoglobin and oxyhaemoglobin can then be quantified using MRI, such that the proton relaxation time measured at a certain voxel is dependent on the amount of oxygen being carried by blood within that voxel.

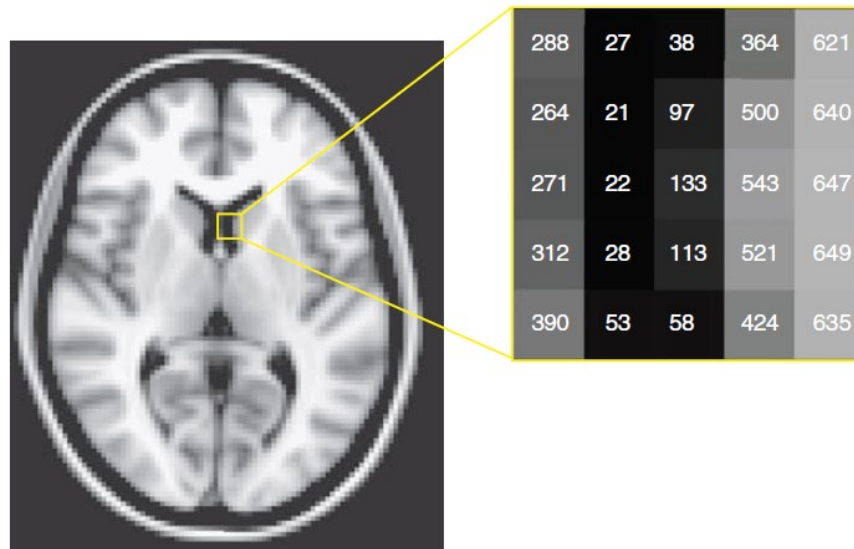


Figure 2-1. MRI image representation. The values of the pixels in the left image correspond to numbers derived from MRI measurement. A subsection of the image with those numbers is given on the right. The image presented here is a 2-dimensional sagittal plane slice from a whole-brain image.

This is interesting from a neuroscientific viewpoint since local increases in neural activity lead to an increase in glucose metabolism in the neurons, which in turn leads to an increase in oxygen consumption (Hyder et al., 1997). Therefore, when neurons become active and oxygen levels increase locally, the measurable BOLD signal is in effect a proxy measure for the activity of neurons in the area, unlocking a method for *in vivo* measurement of neural activity. However, the relationship between neural activity and the subsequent BOLD response is not trivial, and the precise mechanisms are still a subject of debate (e.g. Logothetis & Wandell, 2004; Ekstrom, 2010).

If neurons are stimulated for only a brief period of time, the subsequent signal elicited is described by a haemodynamic response function (HRF) similar to that shown in Figure 2-2. Immediately after the onset of stimulation is a brief period (around 1s) where the BOLD signal decreases. This is followed by an increase in BOLD which reaches a peak around 5s after the activity onset. This oversupply of oxygen to the area of neural activity results in a decrease of deoxyhaemoglobin in the area which is measurable by fMRI. In the last phase

of the HRF, BOLD response undershoots the original baseline before levelling out after around 20s of stimulus onset.

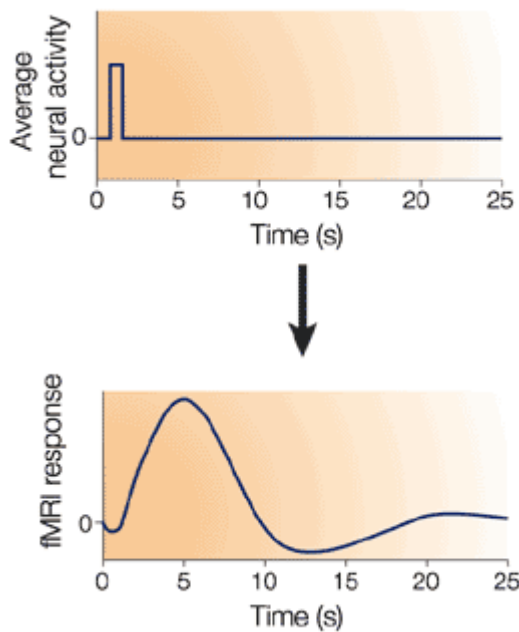


Figure 2-2. The haemodynamic response (that measured using fMRI) induced by a short burst of neural activity (from Heeger & Rees, 2002).

2.1.2 FMRI data analysis

A common step in analysing fMRI data collected during an experiment, and one used throughout this thesis as a basis for more complex methods, is to infer the contribution of experimentally set variables (e.g. whether stimulation is presented at a certain position of the visual field) to the recorded voxel-level BOLD response. This is achieved by framing the BOLD time-series recorded at a voxel as a linear combination of experimental variables (or *regressors*) – the contribution of each regressor to the BOLD response is then estimated by fitting a general linear model (GLM).

2.1.2.1 The general linear model

The general linear model (GLM) for a given voxel's recorded BOLD time-series can be written (in matrix notation) as

$$\mathbf{Y} = \mathbf{X}\beta + e$$

where \mathbf{Y} is a column vector of BOLD responses measured from a single voxel in an fMRI time-series (i.e. each element of the vector is the BOLD response at a given time t); \mathbf{X} is the *design matrix* of the experiment, where each column represents the state of some experimental variable; e is an error vector, where errors are assumed to be independent and identically distributed across the time-series. The term β is a column vector of *effect* parameters, each element represents a scaling factor corresponding to a given regressor in the design matrix. As such, the values of β indicate the extent to which a given regressor drives the BOLD response of a voxel.

The design matrix \mathbf{X} can be composed of partitions $[\mathbf{G} \ \mathbf{N}]$ where \mathbf{G} corresponds to regressors of interest, such as experimental manipulations which are hypothesised to induce changes in neuronal activity, and \mathbf{N} corresponds to nuisance regressors (which may include estimates of subject motion calculated from the volume time-series). The columns of \mathbf{G} may be composed of indicator variables, in the range $[0, 1]$, which correspond to whether a certain condition is currently active (e.g. whether visual stimulation is present), or can represent continuous variables such as subject motion in the sagittal plane. Regressors can also be defined instantaneously at a time t using the Dirac delta function, or can be set across a span of time, t to $t + T$; when such a scheme is used for an indicator variable, the regressor is commonly termed a 'box-car' regressor due to its shape. The final design matrix \mathbf{X} is arrived at after convolving each column

with a HRF, which can either be derived experimentally per subject or the ‘canonical’ HRF which is standard across subjects (Frackowiak et al., 2004).

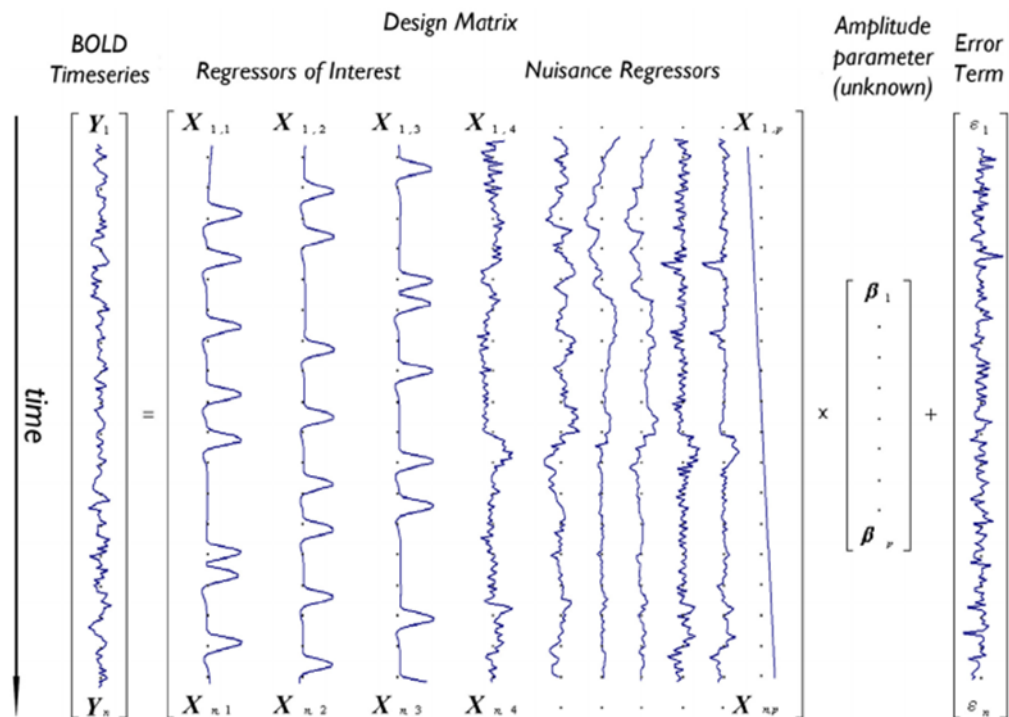


Figure 2-3. Example GLM configuration for a voxel with recorded BOLD time-series Y , 10 regressors X (3 of interest, 7 nuisance variables corresponding to 6 motion estimates and 1 linear drift) each of unknown amplitude β (from Monti, 2011).

The GLM is then fitted using a modified ordinary least squares procedure to find the vector β , such that each element represents the *effect* of each regressor on the voxels’ BOLD time-series. A regressor with an associated β value of 0 would be interpreted as having no effect, either excitatory or inhibitory, on neural activity. In practice, this model estimation procedure is calculated in parallel across all recorded voxels in the volume: a volume of voxels is thus outputted for each column of the design matrix where each voxel value is the associated β value. At this point, one can produce statistical parametric maps (SPM; Friston et al., 1995) using the β volumes to report areas of the brain which are active under certain conditions at a statistically significant difference. However, in the work presented here in Chapters 3 and 4, our use of the traditional GLM approach to modelling fMRI stops with β values, which we use

as measures of activity in each voxel associated with a given stimulus condition. This data can then be used in more complex methods such as multivariate pattern analysis (MVPA; Haynes & Rees, 2005; Kamitani & Tong, 2005) or voxel tuning functions (Serences et al., 2009) to richly characterise the responses of voxels across stimulus variations.

2.1.2.2 Retinotopic mapping

Much of the work in this thesis is concerned with the response of neurons in early visual cortex to differing visual stimuli. We are therefore concerned primarily with the BOLD response of voxels which carry information regarding these neurons, and would like to isolate the activity of these voxel for further analysis. Luckily for us, the visual cortex is *retinotopically organised*, meaning that two relatively proximal positions in the visual field are projected through the retina and optical nerve to relatively proximal patches in visual cortex. Furthermore, each point in the visual field is actually represented many times across distinct regions of the cortical surface, and the enclosed cortical areas representing the full field coincide with the anatomical boundaries of visual areas (e.g. V1, V2, MT etc.). Therefore, by recording activity to certain types of visual stimulus it is possible to isolate the regions of cortex which correspond to these areas through a process called retinotopic mapping.

In Chapters 3 and 4, retinotopic maps, i.e. the correspondences of stimulus position in the visual field to neurons in visual cortex, were extracted using the recently developed population receptive field (pRF) method (Dumoulin & Wandell, 2008). This method estimates an explicit receptive field model for each voxel; that is, the area of the visual field where the voxel is responsive to visual stimulation. Each pRF model is defined as a 2D Gaussian in the visual field with

parameters x_0 , y_0 , and θ ; where (x_0, y_0) represents the centre of the Gaussian, and θ represents its spread (see Figure 2-4 for model fitting details). The value of this procedure is that the estimated voxel-level parameters are connected meaningfully to the neuronal parameters, leading to more accurate mappings (Dumoulin & Wandell, 2008; Alvarez et al., 2015).

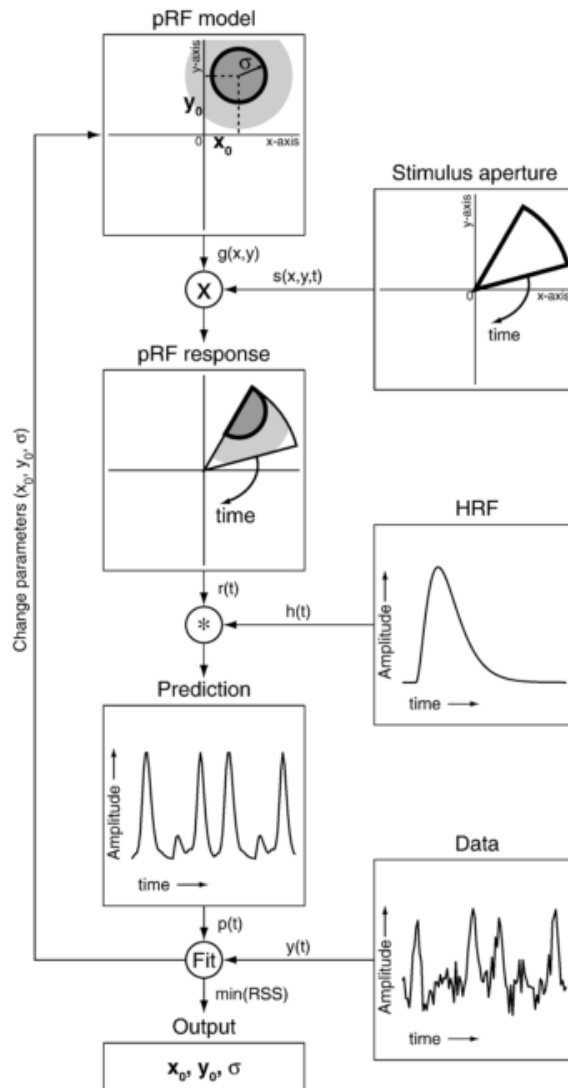


Figure 2-4. Estimation of population receptive fields (pRFs). The pRF approach fits a distributed Gaussian model across a stimulating portion of the field for each voxel. In the model-based analysis shown, a voxel time-series prediction $p(t)$ is calculated as the product of a parameterized model of the underlying neuronal population, $g(x, y)$, and the ring stimulus, $s(x, y, t)$, followed by convolution with the haemodynamic response function (HRF; the relationship between neuronal activity and BOLD signal). The prediction is then compared with the data, and the model parameters optimised with a coarse-to-fine grid search approach.

Using the location estimates of pRF models, early visual areas are manually delineated (see Figure 2-5 for example and process details). In the current work voxels belonging to V1, V2, and V3 were isolated, to allow further analyses across subjects within these regions of interest.

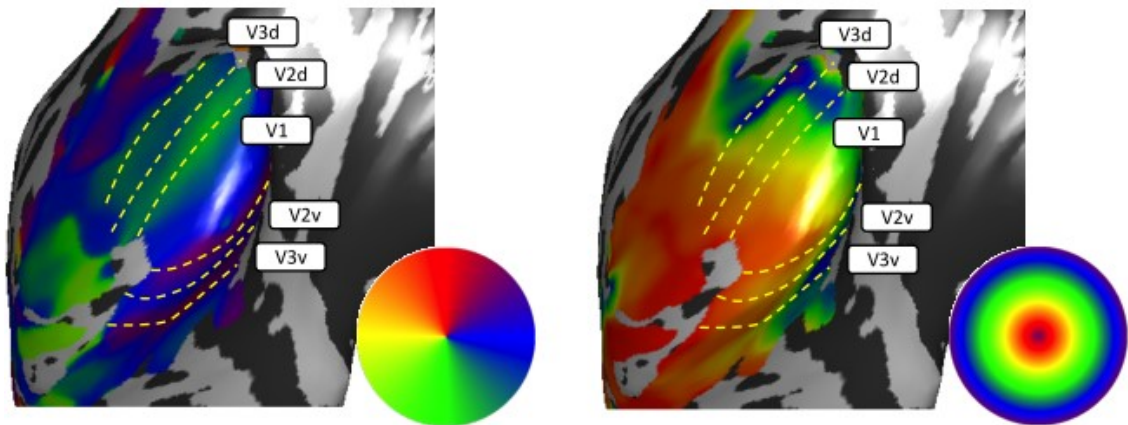


Figure 2-5. Delineation of visual areas. Polar (left) and eccentricity (right) maps are shown projected onto an inflated surface model of visual cortex, with visual area delineations and corresponding labels. Where a map reverses generally coincides with the boundary of a visual area. This is easier to see in the polar angle map on the left: V1 in the ventral-to-dorsal direction maps a clockwise shift in angle (purple to green); after the V2d border, movement across the cortex in the same direction maps an anti-clockwise shift in angle (green to blue); this is again reversed at the V3d boundary. A similar oscillating pattern is seen across the ventral areas.

2.1.2.3 Multivariate pattern analysis

Multivariate pattern analysis (MVPA) methods extract information across sets of voxels such that the information contained in relative activity differences between voxels is utilised. The pooling of such information across sets of voxels can allow differential brain activity between conditions to be observed, even when univariate methods fail to detect effects (e.g. Formisano et al., 2008; Obleser, 2010). There is also a shift of analytical focus associated with MVPA methods: MVPA seeks to *predict* the experimental condition based on voxel responses rather than determine the significance of voxel response differences due to experimental condition. Figure 2-6 gives a simple MVPA demonstration, where the set of voxels is limited to two, and the response amplitudes of each

voxel are given on separate axes. The conditions in Figure 2-6 are well-separated by the classification boundary when responses to both voxels are taken into account, however the Gaussian distributions on the axes indicate that each voxel taken in isolation would be a poor predictor of experimental condition.

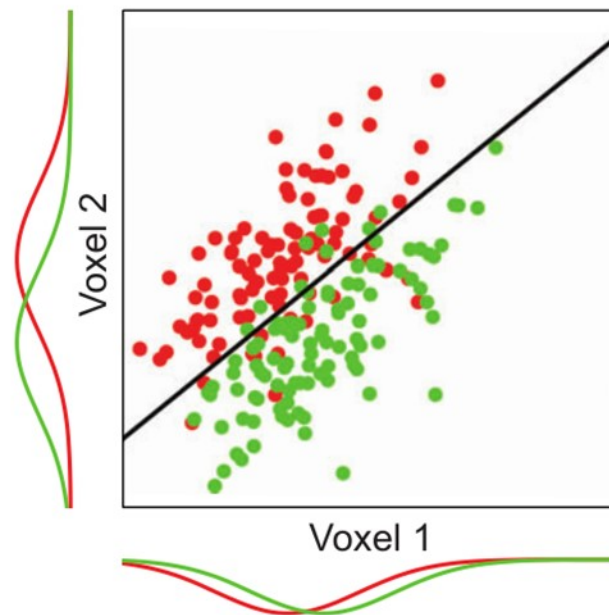


Figure 2-6. Putative responses of two voxels to two experimental conditions. Red circles represent responses to one condition, while green circles represent responses to the other. Colour-coded Gaussian distributions on the voxel axes show the distribution of that voxel's responses to that condition. The line dividing the red and green circles is the maximally separating plane between the conditions, arrived at with a support vector machine (SVM), such that any new data points falling above this line would be classified as red, and any new point falling below would be classified as green

Typically, MVPA progresses in two steps: training and testing. The dataset is first split into training and test sets, for example, if a participant completed 8 scans of an experiment where we are interested in predicting the experimental condition at a given time from brain activity, we could construct a *training set* using data from 7 of the 8 scans, and the test set would comprise data from the remaining scan. Here, the data is structured into exemplars, where each exemplar is a vector of distributed voxels activity in a region of interest with a corresponding condition label. Using the training set, a classification boundary is learned to separate the data generated in one condition versus the other.

This is typically done using a linear support vector machine (SVM; Cortes & Vapnik, 1995), which is used in Chapters 3 and 4 of this thesis, although other approaches are viable (for a review see Misaki et al., 2010). Once the classification boundary is learned, the accuracy of the classifier is measured using the test set – if each pattern of activity in the test set is classified as the correct condition then the classifier has an accuracy of 100%, if only half are correctly classified then the classifier is said to be 50% accurate, and so on. To reduce variance in the accuracy estimate, the accuracy values of such an analysis are averaged across multiple cross-validation folds of the dataset. For the example given above, with 8 completed experimental scans, the data from the first scan is retained as a test set, while the remaining scans' data are used to train the classifier – accuracy of the classifier is then measured on the test set. Then data from the second scan is used as the test set, while the remaining scans are used to train the classifier – again performance is recorded on the test set. This process is repeated for each possible permutation of test and training set (in this example, 8 times), and the accuracies of the classifier in each *fold* are averaged to produce the final accuracy estimate. This accuracy estimate can then be interpreted as a measure of the robust separability of the two conditions in terms of evoked activity across distributed cortical regions.

Most relevant to our work in understanding the effect of perceptual load on the representation of orientation is the finding that orientation-specific information is contained within distributed patterns of brain activity, as illustrated by MVPA. For example, Kamitani and Tong (2005, see Figure 2-7) found that when different orientation gratings are presented to a subject, each grating produces a slightly different response pattern across a selection of voxels in early visual cortex such that classifiers could robustly predict the orientation of a presented oriented grating from brain activity.

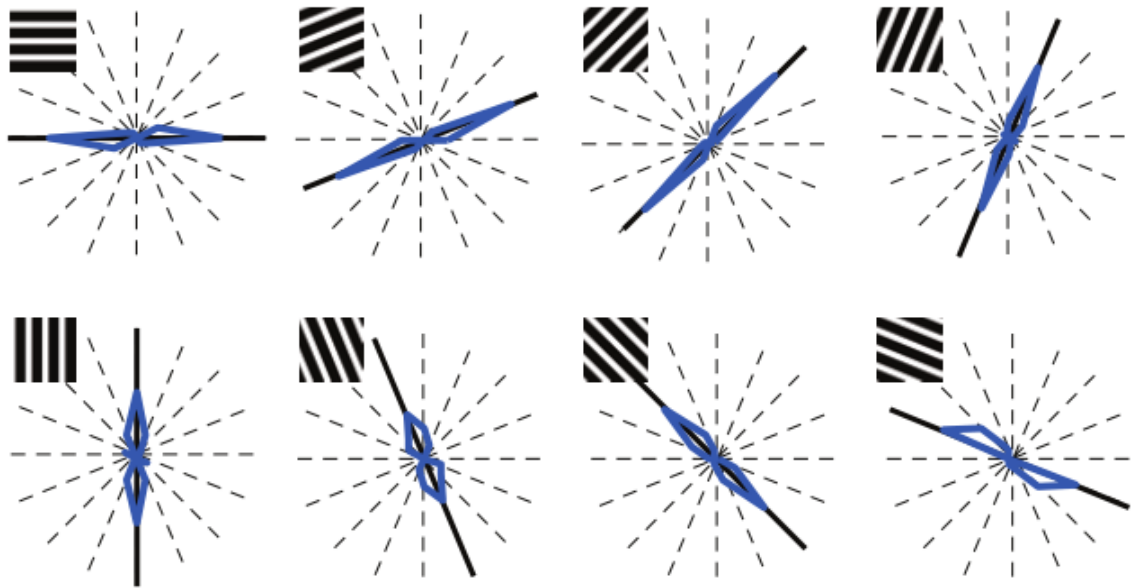


Figure 2-7. Prediction of stimulus orientation with MVPA. Blue curves represent the distribution of stimulus orientations predicted by an ensemble of SVM linear classifiers, while dark lines represent the actual stimulus orientation. The corresponding orientated grating stimuli are shown at the top left of each polar plot (from Kamitani & Tong, 2005).

The essential concept is that neurons within single voxels will have an uneven distribution of orientation preferences, leading to a potential bias in the preference of those voxels. When differently oriented gratings are presented to a subject, each grating produces a slightly different response pattern across the early visual cortex (Kamitani & Tong, 2005). Although the irregularities are very small on a voxel-by-voxel basis (and thus not measurable by traditional univariate analysis), when the distributed pattern of voxel activity is taken together it can provide enough information to accurately classify distinct stimulus orientations using a linear classifier (but see Alink et al., 2013, for a different interpretation of the classification results). Although these methods cannot provide a direct link to the underlying single-cell responses, since other parameters such as the distribution of preferences within voxels and the voxel size contribute to the voxel-level effects, they do reveal the presence of feature specific information and enable the prediction of perceptual states from brain activity (Haynes & Rees, 2006). They have therefore been used to estimate the degradation of representations under varying conditions – for example Kok et al. (2011) found that patterns evoked by oriented gratings were *less* decodable

when the orientation of the gratings was unexpected. The method can therefore be applied to our research question into whether a similar degradation of orientation representation occurs when subjects are under conditions of high perceptual load.

2.1.2.4 Voxel-based tuning functions

VTFs exploit neuronal orientation biases within voxels to construct population response profiles from imaging data; as such, they operate on the same assumptions as MVPA methods. However, whereas classifiers pool feature-selective information from all voxels in a visual area to discriminate activity patterns, VTFs aim to preserve and reveal the tuning properties of individual neural populations within voxels. VTFs extracted from early visual cortex have been shown to be modulated by both feature-based (Serences et al., 2009; see Figure 2-8) and spatial attention (Saproo & Serences, 2010), and so appear ideal to investigate the effect of perceptual load on the low-level encoding of orientation.

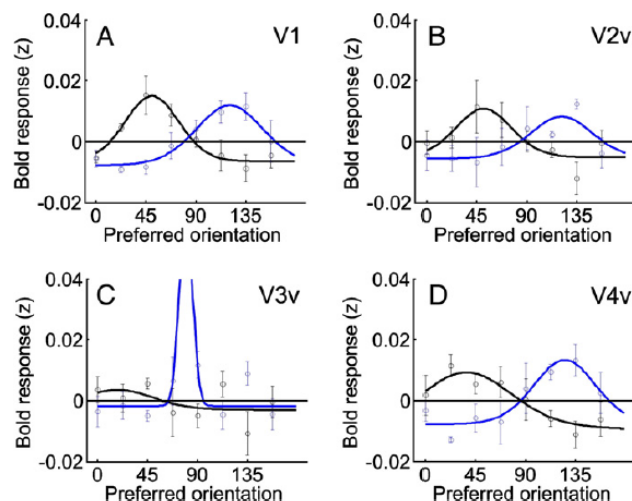


Figure 2-8. Voxel tuning functions obtained when participants were told to attend to either the 45° (black curves) or 135° (blue curves) in an overlapping grating stimulus. The preferred orientation of the VTFs is shifted to the attended orientation, even though both orientations are presented concurrently in both conditions

VTFs can be seen as a characterisation of orientation-dependent response within voxels, and are therefore a non-invasive population-level analogue of traditional single-cell tuning curves. For a given voxel, a BOLD response time-series is collected whilst the subject views a full range of oriented gratings. The responses to these gratings can then be centred on the voxel's preferred orientation (i.e. that which elicits the maximal response among the orientations) and arranged into a tuning curve. The BOLD responses, or fitted GLM parameters, across the orientation range can then be fitted with a circular Gaussian function, as used for characterising single-cell and population-level tuning functions (e.g. Martinez & Trujillo, 2004; Series et al., 2004). This therefore allows the extraction of tuning curve parameters which have been shown to influence the perception of orientations. For example, the width of the tuning curve indicates the precision that the voxel encodes its preferred orientation (Series et al., 2004), while tuning curve amplitude is a major indicator of the information the population is able to convey regarding the stimulus state (Sprague et al., 2015). VTFs can therefore offer a more direct path to the underlying neural encoding of orientation in comparison to MVPA methods, which, although perhaps more powerful in determining *if there exists a difference* between conditions (Saprou & Serences., 2010), obfuscate the population tuning characteristics through linear pooling of responses.

In Chapter 4 we compute VTFs independently under conditions of high and low perceptual load. GLM parameter estimates associated with specific orientations are used as measures of the orientation-specific elicited activity, and VTFs are constructed for each voxel using this data. Similarly to the MVPA approach, a leave-one-scan-out cross-validation procedure is used to construct a voxel's final VTF estimate: using averaged data from $N-1$ scans, the preferred orientation of the voxel is computed; a VTF is then constructed using the data from the other scan, centred on the precomputed preferred orientation. This

procedure is repeated for the N permutations, resulting in N VTFs, which are then averaged to give the overall VTF for that voxel. After averaging across voxels within an ROI, we arrive at a characteristic tuning curve for that ROI – parameters of fitted circular Gaussian models can then be compared across conditions to ascertain whether the experimental manipulation affected low-level orientation encoding and subsequent perception.

Although it would be possible to use independent scans without a load task to train orientation classifiers (in MVPA) or compute voxel orientation preference (in VTF analysis), the cross-validation approach used here compares voxel patterns or activities *between* conditions. Therefore, any change in accuracy or tuning reflects a change in the fidelity of activation; in fact training using ‘no-load’ scans condition may bias the results towards the low-load condition, given that the attentional state with no primary task is more similar to the low load condition than high.

2.2 Computer vision and machine learning

The approach used in this thesis to produce a predictive model of perceptual load is rooted in computer vision and machine learning techniques. At the base of this is to characterise perceptual load as an attribute of the driving scene; we can then associate values of this attribute to segments of driving footage collected from a driver's point-of-view during urban driving. Once a dataset of video segments with associated load values is obtained, the modelling task becomes a regression problem between video segments and load values. In the following sections I describe the methods used to: 1) obtain consistent estimates of the perceptual load attribute from combined judgements of many

human annotators, 2) represent video segments with compact semantically informative descriptors, and 3) map those descriptors to perceptual load values.

2.2.1 Obtaining ground-truth perceptual load values

Subjective estimates of attributes may be obtained by participants viewing stimuli and assigning an absolute attribute value to each viewed stimulus (e.g. from a range of 1-7). This approach, while simple, has drawbacks: the estimates obtained are susceptible to within-subject baseline and variance shifts across time, and have been shown to produce less accurate labellings relative to the pairwise comparison method in fields closely related to the current topic (e.g. medical image assessment: Phelps et al. 2015). More consistent attribute estimates are obtained through participants making *relative* judgements, in a pairwise comparison design (e.g. in estimating the 'shininess' of shoes, Kovashka et al., 2012; and the attractiveness of faces, Donahue & Grauman, 2011). Therefore, in our method, pairs of driving videos are viewed by annotators, whereupon they indicate which video represented the driving situation of highest perceptual load.

To transform a number of relative pairwise comparisons into a continuous measure of load we make use of the TrueSkill algorithm (Herbrich et al., 2006), a method initially developed for calculating the relative skill levels of players in competitor-versus-competitor games such as chess. It is adopted here as it has shown state-of-the art performance in related domains (Chen et al., 2013). In our context, as pairwise comparisons are completed between video pairs, the TrueSkill algorithm maintains estimates of each video's perceptual load value, which are updated as more comparisons are completed. In TrueSkill the perceptual load of each video is represented as a Gaussian distribution, $N(\mu, \sigma)$,

where μ represents the current estimate of the perceptual load, and σ represents the algorithm's current uncertainty regarding that estimate. After each comparison, the load distributions are adjusted. This proceeds such that a currently lowly rated video being judged as higher load than a currently highly rated video results in a large load distribution shift. However, the opposite, expected, result (i.e. a currently high load video being deemed higher load than a low load video) would result in a very small shift of estimated load values. In our implementation, values for each video are initialised, before any comparisons have been made, at $\mu = 25$ and $\sigma = 8.33$ (following Herbrich et al., 2006). After a sufficient number of comparisons, ratings become stable; this occurs at approximately 30 to 40 comparisons per stimulus in most applications. The μ of a video's load distribution is then taken as the ground-truth perceptual load value for that video, resulting in a dataset of video and load value pairs suitable for regression analysis.

2.2.2 Extracting semantics from imagery

A machine-learning approach to extracting semantics from imagery has recently emerged, driven by the availability of increased computational power and large curated datasets from which to learn. In general, this methodology requires formulating the semantic extraction task as classification (i.e. predicting the category of some object in an image or video) or regression (i.e. predicting the value of some attribute of the image/video), then collecting a ground-truth dataset of labelled examples, and learning a mapping using these known exemplars. The approach is therefore suitable for our problem of estimating perceptual load, given the creation of a large ground-truth dataset of driving videos and associated perceptual load values.

The method is long-established and has proven very successful for deriving semantic information, such as object or character identity, from *static* images. There exist two general pipelines for such an analysis. In the traditional approach, images undergo a first feature extraction stage, where informative lower dimensional image descriptors are calculated from the image information, before a mapping is learned between these descriptor vectors and the associated image category or attribute value. Popular descriptors of this type include the histogram of oriented gradients (HOG; used for human detection by Dalal and Triggs, 2005), the scale-invariant feature transform (SIFT; developed for generic object recognition by Lowe, 1999; 2004), and texon forests (developed for semantic segmentation by Shotton et al., 2008). Recently however, a new approach to such tasks has emerged, obtaining state-of-the-art performance in several of these domains. In this approach, termed deep-learning or feature-learning, the intermediate representation by image descriptors is removed as a stand-alone process, with the classifier or regressor learning to map directly from image pixels to the target value. In the image domain, convolutional neural networks (CNNs) have recently shown best performance on popular image understanding datasets, for example Zheng et al. (2014) proposed a neural network combining a CNN with a recurrent conditional random field network for pixel-level semantic segmentation, producing state-of-the-art performance on the challenging *PASCAL VOC 2012* segmentation dataset. Similar strides have been taken in object recognition, with CNN-based network architectures producing the current best results on many datasets (e.g. the ILSVRC2012 dataset, He et al., 2015; the CIFAR-100 dataset, Clevert et al., 2015).



Figure 2-9. Sample video frames with associated action categories from the UFC-101 (top row) and Hollywood2 (bottom row) action recognition datasets.

However, there are greater challenges in understanding *video*: the very high dimensional nature of the data (a video effectively being 30 images captured per second) combined with the need to incorporate and leverage temporal information result in a much more difficult problem. As such, hand-tuned feature extraction pipelines are still among the best, although recent applications of CNN-type network architectures to deep-learn features are approaching similar performance levels. In the video analysis literature, much of the advancement is gauged using a system's ability to classify human actions in video, for example 'picking up a telephone' or 'getting out of a car'. Commonly used datasets used in the literature to gauge performance are Hollywood2 (Marszałek et al., 2009) and UFC-101 (Soomro et al., 2012) datasets, where the current best classification performances are obtained by the traditional descriptor-extraction methodology of improved dense trajectories (IDT; Wang et al., 2013; 2015) and the deeply-learned 3-dimensional CNN architecture of Du Tran (C3D; 2014). Therefore in Chapter 5 we implement IDT and C3D to extract useful visual information from driving videos, and describe the methods in detail here.

2.2.2.1 Improved dense trajectories

Improved dense trajectories (IDT Wang et al., 2013; 2015) is a video representation based on the extraction of appearance descriptors around interest points tracked through time. Throughout our work we use default

parameter values provided by Wang et al. (2015), as they have shown robust performance across many datasets.

Trajectories are sampled densely and tracked using optical flow images for a maximum of 15 frames (to alleviate effects of interest point drift). Sampled trajectories are firstly removed if they are in an area with little texture information or variation (e.g. a monochrome wall) as it is impossible to track a point in an area without structure. The optical flow images used for subsequent tracking are in fact altered from the original estimates supplied by Farneback's algorithm (2003) through computing homographies between successive frames using RANSAC (Vincent & Laganière, 2001) and removing flow consistent with camera motion. This makes the descriptor more sensitive to differences in distinct object motion between videos, and this which sets *improved* dense trajectories (Wang et al., 2013; 2105) apart from the original dense trajectories configuration of Wang et al. (2011).

The first stage after a trajectory is extracted from the video is to describe it using fundamental visual descriptors (see Figure 2-10 for a visual representation of this process). This same process is repeated for each trajectory extracted in a video (which can be upwards of 100,000). The shape of trajectories through the video cube is used as a descriptor of global motion patterns and is encoded as a sequence of 2D displacement vectors along the time-course of the trajectory, $\{(x_0, y_0), (x_1 - x_0, y_1 - y_0), \dots, (x_n - x_{n-1}, y_n - y_{n-1})\}$, resulting in a 30D trajectory displacement (TD) descriptor for each 15 frame trajectory. To embed rich motion and appearance information a space-time volume is aligned with each trajectory; the dimension of the volume being 32 x 32 pixels spatially and 15 frames temporally. Each spatio-temporal volume is divided into a regular grid structure of spatio-temporal cells; each volume being separated into 2 cells horizontally, 2 vertically and 3 temporally, resulting

in a total of $2 \times 2 \times 3 = 12$ cells per trajectory. Histogram of oriented gradients (HOG), histogram of optical flow (HOF), and motion boundary histogram (MBH) descriptors are subsequently extracted in each cell: these descriptors are computed for each frame in a cell and summed across frames. The final descriptor for each trajectory is then taken as the concatenation of these cell descriptors.

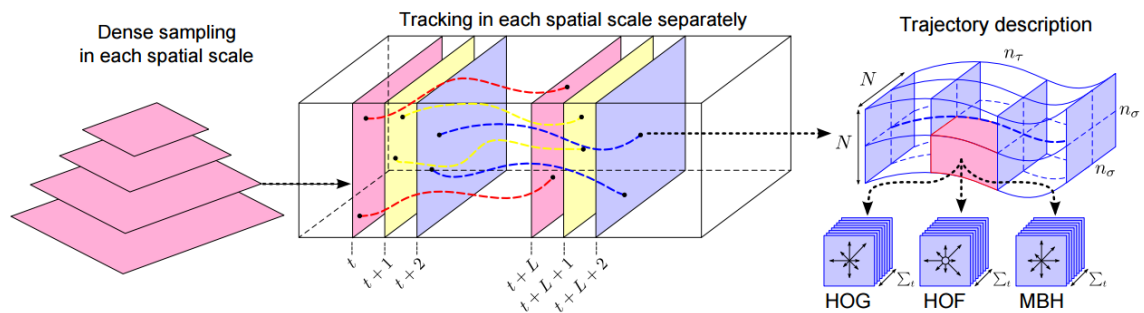


Figure 2-10. Dense trajectories extraction. At the left, basic feature points are densely sampled at a range of spatial scales. After removing points in homogenous areas, each point is tracked for a maximum of 15 frames using optical flow images. Finally, trajectory displacements and descriptors (HOG, HOF, MBH) are computed along each trajectory in a 32×32 pixel neighbourhood, which is divided into $2 \times 2 \times 3$ (x-axis, y-axis, frames) cells.

Gradient and optical flow orientation in the HOG and HOF descriptors is quantised into 8 orientation bins (and an additional 'zero' bin for the HOF descriptor to encode static objects) while histogram entries are weighted by gradient or flow magnitude; each subsequently normalised by its L_2 norm. The final dimension of the HOG feature for a single trajectory is thus 2 (horizontal cells) \times 2 (vertical cells) \times 3 (temporal cells) \times 8 (orientation bins) for a size of 96D, while the HOF feature for a trajectory is of dimension $2 \times 2 \times 3 \times 9 = 108$ D. MBH descriptors are calculated separately for the x- and y-axis components of optical flow images, resulting in distinct MBHx and MBHy descriptors each of dimension $2 \times 2 \times 3 \times 8 = 96$ D. Similarly to HOG, MBHx and MBHy histogram entries were weighted by gradient magnitude and subsequently normalised using the L_2 norm.

From each video clip then, a certain (but variable) number of trajectories and associated descriptors are extracted. To transform this into a fixed-length representation for each video, a bag-of-visual-words approach is used. A codebook is learned for each descriptor type (TD, HOG, HOF, MBHx, MBHy) independently, where codebook size is set at 4000 visual words. The codebook for each type is learned from a randomly selected subset of trajectories from the training data (typically around 100,000) using the k -means algorithm. For a given video and descriptor type, each extracted descriptor is assigned to the closest learned codeword in terms of Euclidean distance; this 4000D histogram is then L_1 normalised; resulting in a 4000D representation vector for that descriptor type. This process is repeated for each of the 5 descriptor types (TD, HOG, HOF, MBHx, MBHy), resulting in a 5-channel video representation, each channel being 4000D.

To represent a video as a single feature vector necessary for a regression analysis, these channels are combined with a multichannel kernel. In kernelising a *single* descriptor vector (e.g. the HOG channel of IDT), the descriptor is transformed into a vector of similarities to the other descriptor vectors in the data set, thus turning the dimensionality of this vector from 4000D to ND (i.e. the size of the data set). Depending on the type of kernel function used in this step, nonlinear mappings are achievable between descriptor and dependent variable. Moreover, once each channel descriptor vector has been kernelised in this fashion, a weighted sum can be taken across channels to give a final ND representation. The kernelisation of the multi-channel training set features can then be defined as

$$K(x_i, x_j) = \sum_c k^c(x_i^c, x_j^c)/A^c$$

where $k^c(x_i^c, x_j^c)$ is the similarity between exemplars x_i and x_j with respect to the c -th channel using kernel function k^c , and A^c is the mean value of similarities for the c -th channel. The choice of kernel function k^c is a user-defined parameter, and is chosen dependent on the data type and the expected form of the mapping between features and dependent variable. Given the data in each IDT representation channel is a discrete histogram, a suitable measure of the similarity between pairs of exemplars is given by the χ^2 kernel, as suggested by Wang et al. (2011; 2013; 2015):

$$k^{\chi^2}(x_i, x_j) = \exp\left(-\gamma \sum \frac{(x_i - x_j)^2}{x_i + x_j}\right)$$

where γ is a free parameter which parameterises the width of the kernel.

2.2.2.2 The 3D convolutional network

A different approach to understanding images and video is given by the convolutional neural network architecture (CNN; LeCun & Bengio, 1995; Krizhevsky et al., 2012). In contrast to IDT and related methods, here an end-to-end, pixel-to-label, neural network architecture is implemented which does not require intermediate descriptors to be explicitly extracted; rather, intermediate representations are learned automatically in the hidden layers of the network through the backpropagation of classification or regression errors. This approach is sometimes termed *feature-learning* or *deep-learning* as the entire classification or regression parameter set is learned from labelled input-output pairs.

Inspired by Hubel and Wiesel's (1960; 1968) model of feature sensitive cells in visual cortex, the early layers of a CNN are constructed as banks of local filters over the input space which are suited to exploiting the local spatial correlations of natural images. CNNs enforce a local connectivity pattern throughout the layers of a network, such that the input to a neuron in layer m is a combination of neural outputs from a subset of layer $m - 1$; where the output neurons in $m - 1$ have spatially contiguous receptive fields. Much like a simplistic hierarchy of the visual system, the stacking of such layers leads to outputs which are increasingly global across the input space, as the effective receptive field over the input becomes larger; this allows later layers in such a network to contain high-level information regarding the entire input, such as whether a certain object category (e.g. a dog) is present.

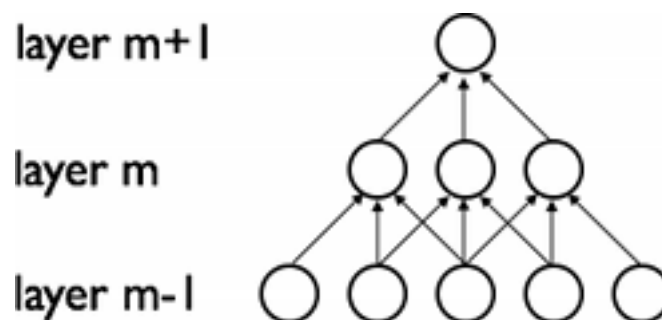


Figure 2-11. 1-dimensional representation of CNN hierarchy and local connectivity. The neuron at layer $m + 1$ is responsive to information change across the whole of layer $m - 1$

Given that CNNs take a high-dimensional pixel representation as input to the lowest layer, and that several layers with associated inter-layer neural connections are implemented in standard architectures, the number of learnable parameters (i.e. the weights of neural connections) in such a model can be astronomical. To improve parameter learning efficiency and aid generalization, a spatially contiguous weight sharing scheme is employed whereby local filters are replicated across the entire input space. This greatly reduces the number of parameters, which is now constrained by the size of the filters rather than the size of the layers. The learning problem is now reduced to learning the weights

of repeated local image filters: for example, one might define filters to be 5 X 5 units in size, and that there are 10 filters to be learned at each layer, across which the output is combined to give the output of the layer.

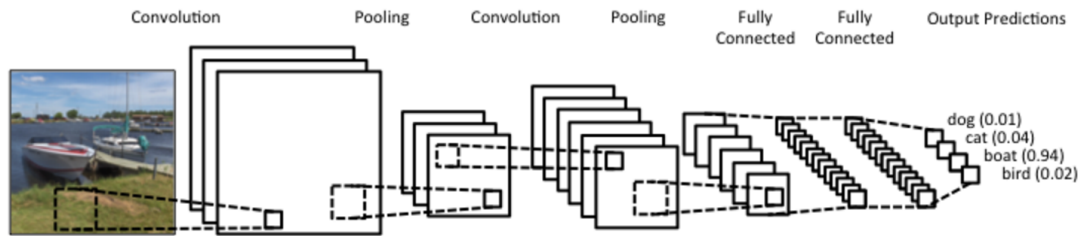


Figure 2-12. Simplified LeNet-5 (LeCun, 1998) CNN architecture set up to recognise an object in an image. In transforming an image through the early layers, filter convolutions are combined with pooling operations (i.e. averaging a contiguous region of input) to reduce the dimensionality of the image to feed as input to fully connected layers. These vector representations are then used as input to a multinomial logistic regression layer which outputs class probabilities.

This configuration results in 5 X 5 X 10 parameters (250) to be learned for a given layer, independent of the input size. In a scheme without weight sharing however, this number is multiplied by the potential number of positions a filter of that size could be placed in the input space, which is roughly equal to the number of input neurons given the locality of filters. Given an input size of 100 X 100 (a relatively low resolution image, for example), the number of parameters to learn explodes to 250 X 100 X 100 = 2.5 million for the layer. Such a scheme also reduces the computation of layer outputs to that of convolution, such that the output of layer m is the convolution of outputs of layer $m - 1$ with a filter f , which, given the existence of highly optimised convolution routines, drastically reduces computation time. The process of obtaining a final network output from an image through successive convolutional layers is described in Figure 2-12.

Given the success of training CNNs on image based tasks such as object recognition (e.g. Krizhevsky, 2012) a natural extension is to the domain of video. A video, being essentially a number of images in a sequence is necessarily of higher dimensionality than an image and therefore is problematic

for a naïve CNN learning approach due the associated increase in number of learnable parameters. However, there also exists temporal redundancy in video (e.g. the appearance of an object will not change much frame-to-frame), and therefore the question of efficiently combining information across video time has received attention recently. For example, Karpathy et al. (2014) approached the problem by taking outputs of image-based CNNs on frames sampled regularly throughout a video (e.g. every 4 frames) and fusing the representations across the video, in a variety of fusion schemes. Given the increased number of parameters learned by such a network, they also collected and introduced a massive video classification dataset, the Sports1M, which consists of 1 million YouTube videos of sporting activities, each labelled with the specific sport. The best of Karpathy et al.'s (2014) architectures obtained 42% accuracy in sport classification (across over 400 categories).

A different approach to combining temporal information was introduced by Du Tran et al. (2014). Instead of combining information across multiple static representations, Du Tran and colleagues alter the convolutional filters themselves to incorporate temporal information. By treating the video as voxel cube, they parameterise 3-dimensional convolution filters at the earliest layers (see Figure 2-13 for description of 3-dimensional convolution).

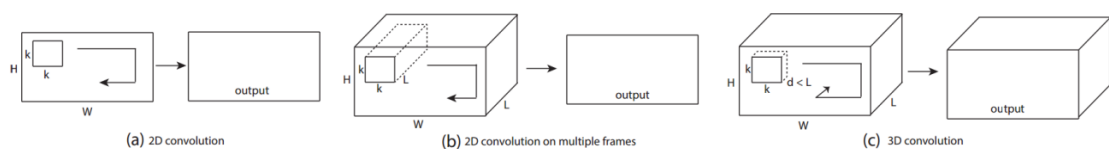


Figure 2-13. 2D and 3D convolution for images and video. a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume and combining the outputs also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal (from Du Tran et al., 2014).

The architecture is therefore analogous to traditional CNNs, however with the addition of a third dimension - time. On the Sports1M dataset, a C3D network

consisting of 8 convolutional layers (see Figure 2-14) achieved state-of-the-art performance of 46% classification accuracy.

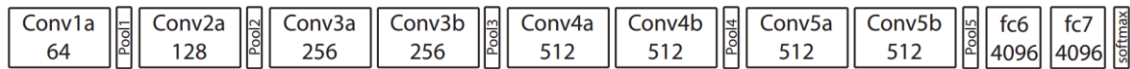


Figure 2-14. C3D architecture. Each Conv layer implements a 3 x 3 x 3 3D convolutional filter, and each pooling operation takes a maximum across 2 x 2 x 2 cells. The number beneath the ‘ConvXy’ text refers to the number of feature maps in that (from Du Tran et al., 2014).

Furthermore, Du Tran et al. (2014) investigated whether the representations of the network, whilst being trained on the Sports1M dataset, would generalise to other action recognition tasks – i.e. were the learned weights generic to understanding spatio-temporal patterns in video. On the UFC-101 dataset, they extracted video representations from the first fully connected layer of the C3D network and trained an SVM classifier to predict actions, achieving state-of-the-art performance. The convolutional filter weights learned using Sports1M videos therefore capture the essence of many motion based activities and concepts in unseen videos, as such we implement C3D in Chapter 5 with the aim of describing the spatio-temporal information present in driving scenarios to predict perceptual load.

2.2.2.3 A novel hybrid descriptor: IDT+C3D

Similarly to the case of action recognition, to estimate perceptual load from video it is necessary to exploit the temporal and spatial context in dynamic natural scenes. Therefore, in Chapter 5, we implement these two (IDT and C3D) state-of-the-art video understanding approaches for the task of estimating perceptual load in driving. Furthermore, we develop a novel hybrid descriptor, termed IDT+C3D, which combines the complementary information captured by IDT and C3D in isolation. IDT is suited to capturing longer-range temporal

information, as its descriptors are collected along 15 frames of videos; however the appearance descriptors it uses are relatively basic (e.g. HOG, HOF). Therefore we inject the richer appearance information captured by C3D (over the shorter time-span of 3 frames) through casting the C3D descriptor as a separate feature channel to be incorporated by a multichannel kernel.

While IDT feature channels are discrete histograms, thus suiting a χ^2 kernelisation, the continuous nature of the C3D representation suggests a radial-basis function (RBF) kernel is more appropriate to characterise the inter-exemplar similarity, such that:

$$k^{RBF}(x_i, x_j) = \exp\left(-\gamma|x_i - x_j|^2\right)$$

where γ refers to a free parameter which parameterises the width of the kernel, and $|\mathbf{z}|$ refers to the L_2 norm of some vector \mathbf{z} . The final kernelised IDT+C3D video representation is therefore a weighted average of IDT channel χ^2 kernel matrices and the C3D RBF kernel matrix. In Chapter 5, we compare the performance of this novel IDT+C3D descriptor against IDT and C3D descriptors in isolation, and furthermore experiment with regression methods, channel weighting schemes, and a baseline linear kernel version of the descriptor where the nonlinear χ^2 and RBF kernels of the IDT and C3D channels are replaced with the simple linear kernel:

$$k^{lin}(x_i, x_j) = x_i \cdot x_j$$

where \cdot refers to the vector-product operator.

2.2.3 Regressing from descriptors to attributes

Once a set of videos is represented by a set of descriptor vectors, a regression model can be fitted between the vectors and associated perceptual load values. Here two regression methods are described, ridge regression and support vector regression, which are implemented and compared in Chapter 5 with regard to performance in the prediction of perceptual load.

2.2.3.1 Ridge regression

Given feature vectors $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$, where n is the number of examples, and $x_i \in R^D$, where D is the dimensionality of the vectors (i.e. 4000 for each IDT channel vector, and 4096 for a C3D vector), a simple linear regression model can be estimated between \mathbf{X} and y , such that

$$\hat{y} = \mathbf{WX}$$

where \hat{y} refers to the model's estimate of the dependent variables y (in our case the ground truth perceptual load values). Fitting such a model entails finding the set of weights \mathbf{W} , which when multiplied by the set of feature vectors \mathbf{X} , produces some minimum error over the estimates \hat{y} and the true values y . In linear regression, this error is the residual sum of squares

$$RSS = (y - \mathbf{WX})^T (y - \mathbf{WX}).$$

Assuming Gaussian observation noise, the maximum likelihood estimator of \mathbf{W} is the global minimum of the RSS with respect to \mathbf{W} , which can be obtained directly using the normal equation:

$$\hat{\mathbf{W}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$$

So far, this is the derivation of ordinary least squares linear regression. However, as we aim to regress not from the original D dimensional feature vectors, but from a kernelised feature space, where $x_i \rightarrow \Phi(x_i)$, and the effective dimensionality of the representation may be infinite (in the case of the radial basis function kernel, for example), it is necessary to introduce a regularisation parameter λ on the magnitude of the weights to avoid overfitting; the resulting model is termed *ridge regression*. This new term penalises the magnitude of the weight vector in the error function, biasing the model against the potential complexity introduced by the extreme effective dimensionality of kernelised feature vectors. The error to be minimised then becomes

$$RSS_{ridge} = RSS + \lambda |\mathbf{W}|^2$$

where $|\cdot|$ again refers to the Euclidean norm. Therefore, when weights \mathbf{W} are large, even if the model has a perfect fit to the training data, the error of the model may be quite large since the complexity implied by large weights is penalised by $\lambda |\mathbf{W}|^2$. The amount of penalisation is governed by λ , which becomes a free parameter of the model, to be tuned experimentally. The optimum estimate of \mathbf{W} then becomes

$$\hat{\mathbf{W}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T y$$

where \mathbf{I} is the $D \times D$ identity matrix. In our setting, the observation matrix \mathbf{X} is not an $N \times D$ feature matrix, but rather a $N \times N$ kernel matrix, \mathbf{K} , where each entry $K_{i,j}$ is the similarity between example i and example j in feature space (i.e. $k(x_i, x_j)$ for some kernel function k). In this case, model fitting proceeds as

above, however \mathbf{K} is now the observation matrix, such that the best estimate of \mathbf{W} becomes

$$\hat{\mathbf{W}} = (\mathbf{K}^T \mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{K}^T \mathbf{y}$$

where \mathbf{I} is an $N \times N$ identity matrix.

2.2.3.2 Support vector regression

Support vector regression (SVR; Cortes & Vapnik, 1995) is a regression method which also parametrises model complexity to avoid overfitting. Similarly to ridge regression, SVR constitutes a linear model $\hat{y} = \mathbf{W}\mathbf{X}$ with a penalty on $|\mathbf{W}|^2$.

However, here the objective to be minimised is framed differently. In the formulation of SVR, $|\mathbf{W}|^2$ is minimised directly subject to the constraint

$$|y - \mathbf{W}\mathbf{X}| \leq \varepsilon$$

where ε is a predefined constant. The tacit assumption here is that a linear function f exists which approximates y with ε precision for all examples.

However, this is usually not a justified or desirable assumption, as we may like to allow some errors to enable a simple model. Slack variable may therefore be introduced δ_1 and δ_2 and the optimisation statement is now to minimise

$$|\mathbf{W}|^2 + C \sum_n \delta_1 + \delta_2$$

subject to

$$y - \mathbf{WX} \leq \varepsilon + \delta_1, \text{ and}$$

$$\mathbf{WX} - y \leq \varepsilon + \delta_2$$

The parameter C now determines the complexity of the fitted model, determining the trade-off between complexity and the extent to which deviations larger than ε are tolerated. C is therefore quite similar to the λ parameter in ridge regression, and is also tuned experimentally. This new minimisation problem is solved using a quadratic programming method (for details see Vanderbei, 1999), to arrive at feature weights \mathbf{W} which define the mapping from feature vectors \mathbf{X} to load values y . Note that in our context again, the feature vectors are in fact kernel-based similarity vectors, \mathbf{K} , which does not change the SVR fitting procedure.

2.2.4 Tuning model hyperparameters

For each feature channel fusion and regression pipeline there exist parameters of the system (termed hyperparameters) which affect its performance, for example the λ parameter in ridge regression dictates the extent of regularisation in the model, while the width of an RBF kernel, γ , parameterises the influence of each data point in the model, such that a lower γ specifies a larger field of influence. In practice, these model-level parameters can heavily influence the performance of a model configuration, and the extent of this influence is often dependent on the dataset of interest. In the case of the original IDT formulation in combination with SVR, for example, there are 6 tunable hyperparameters of the system: the complexity penalty C , and the width of each of the 5 channels' χ^2 kernel; a procedure is therefore required to estimate good values of these hyperparameters.

A common method for estimating hyperparameters is 'grid-search', where the cross-validation loss on the training set is calculated across predefined parameter combinations. For example, in a model with 2 hyperparameters, $f(x; \theta_1, \theta_2)$, we may define points in the parameter space for the model performance to be evaluated at, such that the domains of the parameters are $\theta_1 \in \{0,1,2,3\}$ and $\theta_2 \in \{0,1,2,3\}$. The grid-search method exhaustively evaluates the model using each possible combination of these parameters: in this toy example $4 \times 4 = 16$ model evaluations are required (e.g. $f(x; 0,0), f(x; 0,1), \dots, f(x; 3,3)$). Such an approach becomes untenable with increasing numbers of parameters however, as the number of model evaluations, which may be computationally expensive, increases exponentially with parameter size. Therefore, we employ a sequential model based optimisation (SMBO) method based on a tree of Parzen estimators (TPE; Bergstra et al., 2011), which has been shown to outperform grid-search and random search for parameter tuning. This SMBO method iteratively approximates the response surface of model performance with regard to its hyperparameters. At each algorithm iteration, the response surface is calculated, and the next configuration of hyperparameters is sampled. This sampling is biased towards parameter configurations which are predicted to give the highest expected improvement (EI; Jones, 2011) in model performance. The initial sampling distribution of each tuneable hyperparameter is set *a priori*, along with the number of iterations for the algorithm.

3 The effect of perceptual load on representations of orientation

3.1 Chapter Introduction

As reviewed in the general introduction, it is well established that the perceptual load of a task determines the effect of attention on visual perception (see Lavie, 2005). Indeed, high perceptual load of a task is a major determinant of *inattentional blindness* - a phenomenon whereby observers fail to perceive stimuli presented in plain view (Mack & Rock, 1998; Carmel et al., 2011; Cartwright-Finch & Lavie, 2007). The mechanism of this attenuation of perceptual processing has been investigated in several functional magnetic resonance imaging (fMRI) experiments; finding that visuo-cortical neural activity induced by task-irrelevant stimuli is suppressed when the task at hand exhausts our limited perceptual processing capacity (e.g. Rees et al., 1997; Schwartz et al., 2005; Yi et al., 2004); and furthermore that this suppressive effect extends across much of the visual system, from the lateral geniculate nucleus (O'Connor et al., 2002) to areas responsible for recognition of complex shapes and scenes (e.g. Pinsk, Doniger, & Kastner, 2004). Neuroimaging studies conducted so far suggest a simple and appealing explanation of perceptual deficits due to high load, then: when perceptual resources are exhausted by an attentionally demanding task, there is reduced activity in response to task-irrelevant stimuli leading to reduced contributions to the neural representations involved perception, detection, and recognition.

Reduced neural signal may also be accompanied by increased noise, however. At the level of neural populations which are known to be tuned to particular features, for example specific orientations (e.g. Serences et al., 2009) or motion

directions (Rees, Friston, & Koch, 2000; Bartels, Logothetis, & Moutoussis, 2008), this effect of load may result in broadening of feature-specific tuning curves. Neural population response to an oriented grating stimulus, for example, would then be less precise, resulting in less clearly separable patterns of activity induced by differing orientations. In this chapter, I investigate whether the behaviourally established perceptual deficits associated with increased task load may be attributed, at least in part, to changes in feature-specific representations in primary and early visual cortex.

The experiments in this chapter concentrate on the representation of orientation; as a fundamental building block of visual perception, a degradation in the representation of orientation could hinder the formation of coherent percepts throughout higher levels of the visual hierarchy, resulting in diminished ability to detect, localise, and recognise behaviourally relevant information. For example, diminished orientation representations could contribute to load-induced response-competition effects in the Eriksen flanker task (e.g. Lavie & Cox, 1997), which rely on the discrimination of the letter's 'X' and 'N' in the target and in the flanker. If viewed as a horse-race model of stimuli competing to reach perception, the flanker is less likely to interfere with the target identity when its representation is noisier, as would be the case with degraded orientation representations in early vision. The hypothesised modulation of orientation-specific representational content by perceptual load is measured using recently developed multivariate pattern analysis (MVPA) methods (e.g. Haynes & Rees, 2006; Kamitani & Tong, 2005). Using these methods it is possible to estimate the representational content of neural populations in visuo-cortical areas under conditions of low and high perceptual load - if high perceptual load indeed degrades low-level cortical responses to oriented stimuli in a manner similar to that observed psychophysically (e.g. Stolte et al., 2014), then the accuracy of inferences regarding the orientation of an oriented stimulus

from the corresponding induced pattern of BOLD response across a visual area will be reduced.

3.2 Experiment 1

The purpose of this behavioural experiment was to confirm the modulation of orientation perception by perceptual load using an experimental design and associated stimuli suitable for a fMRI experiment. Stolte et al. (2014) psychophysically constructed orientation tuning curves through varying the orientation offset between a large noise mask with that of a small vertically oriented grating and measuring subjects' accuracy at detecting the grating. They were then able to compare parameters of the tuning curves across perceptual load condition to establish that perceptual load acts to reduce amplitude and increase bandwidth of tuning curves; in essence, reducing orientation selectivity at the perceptual level. However, the design of Stolte et al.'s study is not suitable for direct replication in an fMRI experiment as, firstly, their load task employs different stimuli in the low vs. high load conditions, which would affect visuo-cortical response independent of load; and secondly, their detection task uses noise masks and small oriented gratings which would induce little to no measurable orientation-specific activity in visual cortex. Experiment 1 therefore examines whether the modulation of orientation perception can be replicated with stimuli shown to drive orientation-specific activity in early visual areas in many fMRI experiments (e.g. Haynes & Rees, 2006; Kamitani & Tong, 2005), through the introduction of a novel orientation change detection paradigm. An established perceptual load manipulation (e.g. Schwartz et al., 2005) is employed to ensure no physical differences between stimuli in high and low load conditions, to exclude explanations of potential representational modulations in terms of primary task characteristics.

3.2.1 Methods

Participants

10 participants, 4 of whom were female (aged 19 – 32) participated in the experiment. All were recruited from the Institute of Cognitive Neuroscience subject pool, had normal or corrected-to-normal visual acuity and were naïve as to the purpose of the study. All aspects of the study were in accordance with the local ethics committee at University College London.

Apparatus

All stimuli were created using MatLab (2011a, The MathWorks, Nattick, MA) and Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) and presented on a 21" Monitor (1024 x 768 pixel resolution, 75Hz refresh rate) in a darkened room. Participant viewing distance was maintained at a constant 57cm with a chinrest.

Experimental design

A dual-task paradigm was employed: subjects were instructed to detect changes in the orientation of a flickering oriented grating whilst concurrently responding to a rapid-serial-visual-presentation (RSVP) task at fixation, which could either require low or high levels of load on perceptual processes. The magnitude of the orientation changes was varied, allowing the construction of logistic response curves in each perceptual load condition; parameters of the fitted curves could then be compared across conditions to assess the effect of perceptual load on orientation perception.

For the RSVP task, a rapid continuous stream of crosses subtending 2° of visual angle vertically and 1.2° horizontally was presented at central fixation (250ms duration, with 750ms inter-stimulus intervals). The crosses could either be presented 'upright' or 'inverted' corresponding to the crossbar being offset $\pm 0.25^\circ$ from vertical centre, and their colour was drawn randomly from 5 possible colours: red, blue, green, yellow, and brown. Within an experimental session, participants performed low-load and high-load versions of the RSVP task: the low-load version required a speeded left index-finger key-press for any red coloured cross, regardless of uprightness, while the high-load version required a speeded response for either an upright-yellow cross or an inverted-green cross - meaning participants had to monitor for specific conjunctions of shape and colour throughout the high-load streams (see Figure 3-1).

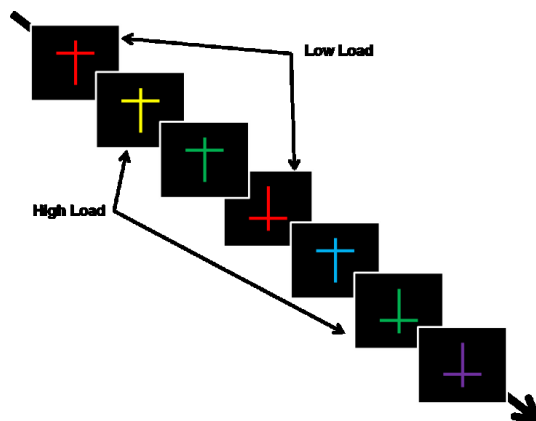


Figure 3-1. The RSVP task used to manipulate load in the experiment, targets in each load condition are highlighted. Note that streams are identical in high and low load, the only difference between conditions being the task instructions

Responses to crosses and associated reaction times were recorded in a 1100ms temporal window beginning 100ms after stimulus onset. In both versions, target crosses requiring a response were presented randomly on 10% of the presentations under the constraint that targets were not presented sequentially, and targets in one condition appeared with equal frequency as

distractors in the other condition to ensure that task instructions were the only discriminating feature between the two perceptual load conditions.

Concurrently with the RSVP task, participants completed a change detection task in the periphery. The task consisted of an oriented grating flickering in synchrony with the presentation of the RSVP crosses (i.e. 250ms duration, 750ms inter-stimulus interval). The default orientation of the grating was 45°, however on 15% of the presentations the grating was rotated away from 45°; the participant was instructed to make a right-hand key-press if this change in orientation was detected. A response was collected in a 2000ms temporal window after the presentation of a rotated grating. A set of 7 orientation displacements were used, which appeared clockwise or counter-clockwise with equal probability. This set of seven orientation changes was decided before the experiment with a short adaptive experiment, so as to probe informative orientation ranges for each participant; a typical range consisted of (1°, 4°, 7°, 10°, 15°, 25°, 35°). There was at least a single trial gap between orientation change trials, allowing the orientation to return to 45°. The full-contrast sinusoidal grating had a spatial frequency of 0.5 cycles/° and subtended 20° visual angle with a central circular aperture, subtending 2.5°, removed for the placement of the load task. The circular edges of the grating were smoothed by convolution with a Gaussian function, so as not to elicit contrary orientation perception, while phase was set randomly on each trial to attenuate a possible motion effect.

Each RSVP stream (and corresponding stream of gratings) consisted of 64 stimulus presentations. Each subject completed 5 blocks, each containing 8 streams, with a short break in between each block. Within each block, load level was counterbalanced in an ABBABAAB or BAABABBA fashion, and this counterbalancing was alternated across participants. Each stream began with a

3s instruction cue consisting of a fixation-dot at centre along with the 2 types of cross targets in the following stream (in low load: upright-red and inverted-red crosses; in high-load: upright-yellow and inverted-green crosses). Then the stream began with the synchronous display of RSVP crosses and orientation gratings.

3.2.2 Results

Load task

Reaction times were significantly slower under high load ($M = 650\text{ms}$) compared to low load ($M = 541\text{ms}$), $t(9) = 3.56$, $p < .01$. Accuracy under high load ($M = 87\%$) was also significantly lower than under low load ($M = 96.6\%$): $t(9) = 3.27$, $p < .01$. These results confirm that the load manipulation was effective.

Orientation change detection

For each participant, change detection accuracies were computed as percentages under each load condition after collapsing across orientation change magnitude.

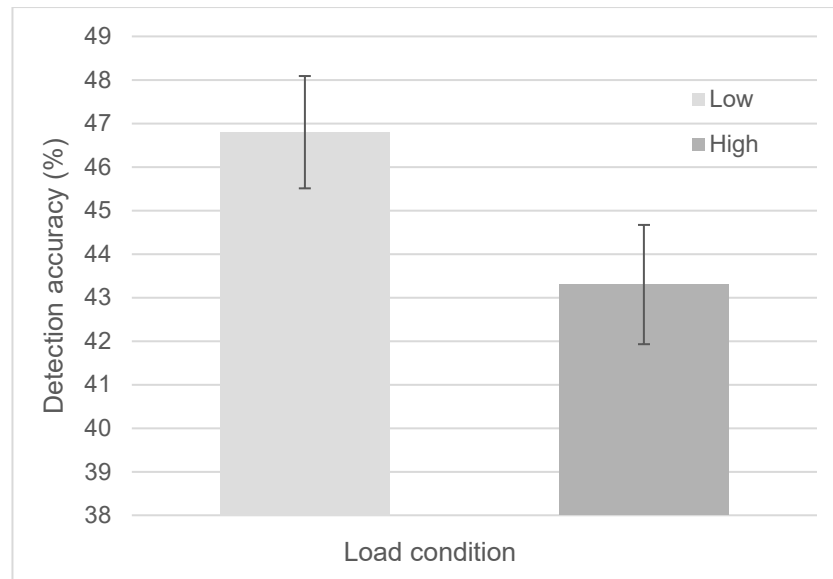


Figure 3-2. Bar graph of orientation change detection accuracy under conditions of low and high perceptual load. Error bars represent \pm SEM across participants

In agreement with our hypothesis, hit rates were reduced under conditions of high perceptual load ($M = 43.38\%$, $SD = 3.87\%$) in comparison to low load ($M = 46.60\%$, $SD = 4.12\%$) in the primary task, a result confirmed as statistically significant by paired t-test, $t(8) = 3.80$, $p = .002$.

3.2.3 Discussion

Primarily, the results of Experiment 1 demonstrate a strong modulation of orientation perception due to the manipulation of perceptual load as change detection was reduced when the central task placed high demands on perceptual processes. One limitation of the results reported here is that only change detection hit rates were collected, leaving open the possibility that the change in hit-rates is due to a criterion shift; a true signal-detection analysis, reporting d' and criteria independently would have been more suited to finding perceptual sensitivity shifts induced by load. The results however are in line with previous findings of reduced detection of unrelated stimuli when under high load, both in the general case (e.g. Carmel et al., 2011; Macdonald & Lavie,

2008) and in the specific case of orientation perception (Stolte et al., 2014). The experimental findings therefore indicate the suitability of both the load task and the oriented grating stimulus for use in fMRI experiments investigating the neural effect of perceptual load on orientation processing in visual cortex.

3.3 Experiment 2

After confirming an effect on the perception of fMRI-suitable oriented grating stimuli due to perceptual load, the source of this effect is examined: can the modulation in part be traced to neural responses in early visual cortex? An fMRI experiment is reported here which utilises MVPA of fMRI data to investigate this possibility.

Previous research in attentional modulation using MVPA (e.g. Kamitani & Tong, 2005) has concentrated on decoding differences between orthogonal orientations. In our case, a similar design would be to present oriented gratings at 0° and 90° under high and low perceptual load, and compare classification accuracy across these two orientations. This approach may lack sensitivity for our question however: it is feasible, and consistent with the psychophysical findings of Stolte et al. (2014), that classification performance would be unchanged between load conditions at extreme orientation differences, with a modulation due to perceptual load becoming apparent only for more distinct orientation discriminations. A larger selection of orientations was probed in order to account for this possibility.

3.3.1 Methods

Participants

Nine adults (aged 19 – 27), 4 of whom were female, participated in the experiment. All were recruited from the Institute of Cognitive Neuroscience subject pool, had normal vision and were naïve as to the purpose of the study. All aspects of the study were in accordance with the local ethics committee at University College London. Each participant completed 15 minutes of training immediately before entering the scanner to get accustomed to the task.

Experimental design

A rapid serial visual presentation (RSVP) task was used to manipulate perceptual load while participants passively viewed oriented sinusoidal gratings in the periphery. The RSVP task stimuli were identical to those used in Experiment 1: crosses were presented for a duration of 250ms, with an inter-stimulus interval of 750ms. Constraints on target placement and target frequency within the RSVP stream were also identical. Left-hand key-press responses to the load task, on a MRI-compatible button box rather than standard keyboard, were recorded up to 1100ms after stimulus onset.

During each scan participants completed a low-load and a high-load stream of the RSVP task. Before the beginning of each stream, an instruction cue was presented for 3s which consisted of a fixation dot at screen-centre along with the two types of RSVP targets to monitor for in the upcoming stream (for low load: upright-red and inverted-red cross; for high load: upright-yellow and inverted-green cross). After each stream was complete, a fixation dot on a mid-gray background was presented for 15s (to measure baseline visual activity in later analysis).

During each stream, an oriented sinusoidal grating was presented in the periphery. The dimensions of the oriented gratings were identical to those in Experiment 1 (subtending 20° in total, a 2.5° central aperture for presenting the RSVP task, random phase etc.), however the grating was flickered at a constant rate of 2Hz, synchronised with the RSVP cross presentations: 250ms duration with a 250ms inter-stimulus interval, resulting in 2 presentations of an oriented grating per presentation of a central cross. The orientation of the grating was changed every 15 seconds, and could take a value from the set $[0^\circ, 10^\circ, 20^\circ, \dots, 90^\circ]$, which was uniformly sampled without replacement at each transition; therefore each orientation was presented once for 15s per load stream. Participants completed 6 experimental scans in total, each lasting 336s, while the load condition was counterbalanced across scans in an ABBABAAB fashion, which was alternated across participants (see Figure 3-2 for visual depiction).

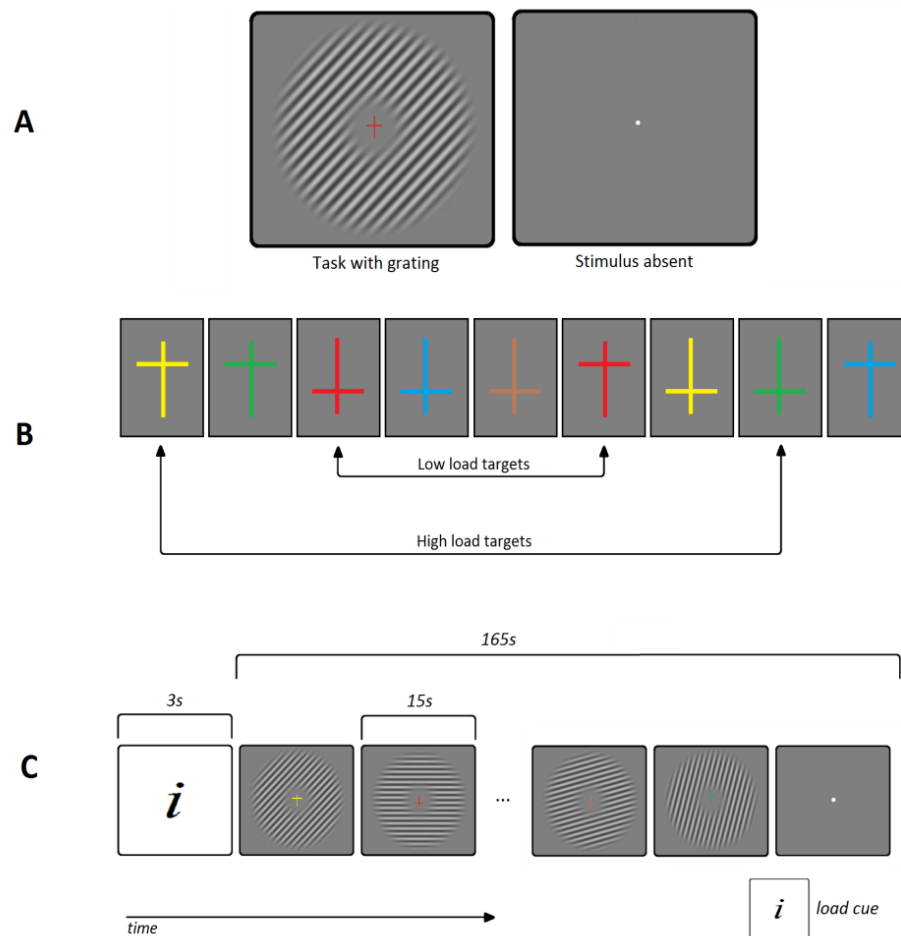


Figure 3-2. Schematic of Experiment 2. A) The two types of display used – peripheral grating and a mid-gray blank display with fixation dot, **B)** an example central RSVP stream with high and low load targets identified, **C)** A representation of a single load condition section of a scan: the orientation of the grating changes every 15s (note that the 2Hz grating flicker is not shown here), while the central RSVP task is sustained throughout. There are two such sections continuously presented in each scan, signalled to the participant by a load cue.

Retinotopic mapping design

For each participant, retinotopic mapping data was obtained with two separate functional scans, recorded the same session as the experimental scans.

Subjects viewed a dynamic, high contrast pseudo-checkerboard carrier pattern which varied continuously in frequency and phase (see Figure 3-3). This carrier pattern was constrained to a disc subtending 9° of visual angle from fixation.

Portions of the carrier pattern were systematically made visible by the application of a combined ring and wedge aperture (see Figure 3-3; Alvarez et

al., 2015). The aperture consisted of a triangular wedge extending from fixation to the edge of the pattern, covering 18° of arc at the carrier disc circumference, which rotated either clockwise or counter-clockwise around fixation, in combination with an expanding or contracting circular ring aperture.

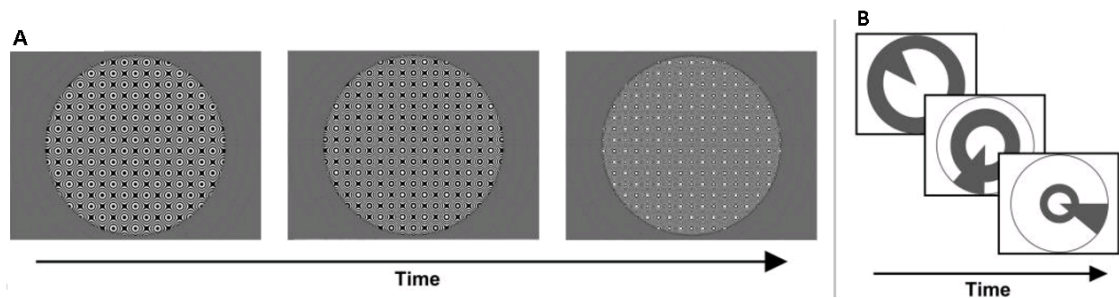


Figure 3-3. Depiction of the retinotopic mapping stimuli. A) The carrier pattern disc, subtending 9° of visual angle radially, and B) the combined ring and wedge aperture which reveals portions of the underlying carrier pattern (figure adapted from Alvarez et al., 2015).

The triangular and ring portions of the aperture were varied at different frequencies: the ring contracted from its outermost position to its most central in 8 steps, where radial size was reduced or increased according to a logarithmic scale such that there was a 50% radius overlap between successive rings; in contrast the accompanying wedge did not vary in size but rotated around fixation, constrained to 6 angular positions covering the entire polar range. Participants completed 2 successive mapping scans, each consisting of 20 cycles of wedge aperture position and 15 cycles of the ring aperture position (i.e. 120 volumes) followed by 24 volumes of uniform mid-gray presentation. One scan presented the apertures in a clockwise/expanding motion configuration, while the other presented in a counter-clockwise/contracting configuration.

fMRI data acquisition

Experimental and retinotopic fMRI scanning was carried out with a 1.5T Siemens Avanto scanner fitted with a 12-channel head-coil. Both experimental and retinotopic scans used the same acquisition parameters. Functional images were acquired using a gradient echo planar imaging (EPI) sequence, manually positioned after a localiser to cover the occipital lobe. fMRI volumes were acquired with the following parameters: flip angle = 90°, bandwidth = 1474 Hz/pixel, TR = 3000 ms, TE = 37 ms, voxel size = 2.3 X 2.3 X 2.3mm, and 36 2.3mm-thick slices were obtained in a descending sequence. A short T₁-weighted MPRAGE image was then acquired. Following this, the 12-channel head-coil was exchanged for a 32-channel head-coil to capture a high-quality T₁-weighted MPRAGE image for detailed cortical surface reconstruction and EPI registration using the following parameters: voxel size = 1 X 1 X 1mm, 176 slices, TR = 2730 ms, TE = 3.57 ms, flip angle = 7°.

Eye tracking

Eye tracking data was recorded with an EyeLink 1000 MRI-compatible tracking system throughout experimental scans. Time-series were obtained of right-eye pupil-position for 5 of the 9 participants at a sampling rate of 500Hz. While technical impediments prevented us from recording eye data for the remaining participants, eye-position was monitored manually by the experimenter in the control room using the EyeLink 1000's video output to ensure central fixation for the duration of the experiment.

3.3.2 Data analysis

fMRI image processing

All functional scans (including both retinotopic mapping scans) were preprocessed using the SPM8 software package in MATLAB (www.fil.ion.ucl.ac.uk/spm). Due to the use of a 12-channel head-coil all scans were individually bias-corrected. After the removal of 4 dummy volumes at the beginning of each scan, 6D affine transformation matrices relative to the initial volume of the scan were extracted for each subsequent volume in the time-series and self-applied, thus realigning each volume with the first in the series. Functional scans were then coregistered with the fast T₁-weighted structural scan. The transformation between the short T₁-weighted anatomical scan and the long T₁-weighted scan, obtained with the 32-channel coil, was then calculated and applied to the functional scans to coregister the functionals with the high-quality anatomical. All functional volumes were finally smoothed with a 6 X 6 X 6mm FWHM Gaussian kernel.

fMRI experiment modelling

Processed experimental scans were modelled as a generalised linear model (GLM) using the SPM8 software package in MATLAB. BOLD time-series for each scan were treated separately but were subject to an identical GLM modelling procedure. The GLM was constructed with boxcar regressors for each distinct orientation presented within each load condition; resulting in 20 orientation regressors per scan (1 for each orientation x load condition combination). Each boxcar regressor was 15s in length, beginning at the stimulus onset of a specific oriented grating and ending when the orientation of the grating changed or the stream ended. A single boxcar regressor for each scan was also included to capture the activity during the 15s mid-gray presentation, and a constant intercept regressor was also included to account for mean BOLD activity. The 6D affine movement estimates, extracted at the preprocessing stage, were also included as continuous regressors in each run.

This resulted in a GLM containing 28 regressors per scan, for a total of 140 regressors across all 5 experimental scans.

Population receptive field (pRF) fitting

pRF mapping was used to exclude voxels which do not convey experimentally relevant information; in our case, for example, voxels not located in the visual system would not contribute meaningful information regarding the perception of oriented gratings. Retinotopic maps, i.e. the correspondences of stimulus position in the visual field to neurons in visual cortex, were extracted using the recently developed population receptive field (pRF) method (Dumoulin & Wandell, 2008) as described in Section 2.1.2.2. Visual areas V1, V2, and V3 were manually delineated for all participants.

Region of interest selection procedure

To ensure that only voxels sensitive to the retinotopic position of the experimental stimuli were selected for further analysis, the experimental scan data were analysed with a separate GLM. Note that selection based on the experimental scans is justified here as the contrast of interest is orthogonal to the experimental hypotheses. The GLM was specified with a boxcar regressor marking the presentation of oriented gratings (i.e. the first 150s of each stream) and then a regressor marking each 14s mid-gray baseline presentation. Following Haynes and Rees (2005), a SPM was calculated from a contrast of the grating regressor vs. the mid-gray regressor, and voxels sorted by t -value; the top 100 voxels by t -value within each visual area, corresponding to the 100 voxels which responded most strongly to the experimental stimuli, were then selected for subsequent analysis.

Multivariate pattern analysis

A linear support vector machine classifier was used to characterise the hypothesised representational degradation associated with increased perceptual load. For determining classification accuracy for each subject in each load condition, a 6-fold cross-validation scheme was used: classifiers were trained on data from 5 of the 6 scans, while performance was computed as the classification accuracy of the data in the 6th scan; this process was repeated 6 times, with data from each scan being used once as the test set. Final accuracies in each condition for each participant were taken as the mean accuracies across folds.

As input to the classifiers, beta values from the fitted general linear model (GLM) specified earlier were used (as described by Misaki et al., 2010). The GLM contained regressors for each 15s orientation presentation within each load stream, resulting in 10 orientation stimulated beta-volumes per run, per load condition. GLM parameter values were extracted for the 100 voxels corresponding to the ROI described earlier (i.e. the most responsive to the presentation of experimental stimuli) within each retinotopically defined visual area. Brain activity to each orientation in each load stream was therefore represented by a 100D feature vector.

In each fold of the cross-validation procedure, data in 5 of the 6 scans served as training data, resulting in a training set consisting of 5 exemplars of each orientation class (i.e. 5 100D vectors per orientation). As SVMs are binary classifiers (i.e. aim to distinguish between two classes of data), and our data has 10 classes (of orientation), a one vs. one multiclass method was used: an ensemble of 45 SVMs was constructed, each trained to distinguish a certain orientation from every other orientation (i.e. all possible pairs of orientations: [0

vs. 10, 0 vs. 20, ... 80 vs. 90]). The test data in a cross-validation fold consisted of a single 100D vector for each orientation - at test time these were fed into each of the 45 classifiers in the SVM bank such that each classifier predicted a class for the sample. The class assigned to the sample was that which elicited the most 'votes' across the classifier ensemble.

3.3.3 Results

3.3.3.1 Behavioural results

RSVP task. Mean detection latencies for targets in the central RSVP task were significantly longer under high load than under low load conditions, 666ms vs. 531ms, $t(8) = 8.33$, $p < .001$. Accuracy under high load ($M = 94.3\%$) was also shown to be significantly lower than under low load ($M = 99\%$): $t(8) = 2.75$, $p < .05$; performance was well above chance in both conditions, indicating that participants understood the task instructions and maintained concentration on the central task. The results therefore suggest that perceptual load was successfully varied by the central RSVP task.

Eye-tracking. Eye position time-series between conditions were compared by calculating the average position and standard deviation of eye position across scans in the experiment. For the 5 participants we were able to collect eye-tracking data from, gaze was highly stable in both perceptual load conditions with a mean offset-from-fixation in low load of 0.37° ($SD = 0.16^\circ$), and a mean offset under high load 0.38° ($SD = 0.08^\circ$), the difference between conditions was not significant, $t(4) = 0.39$, $p = 0.71$. Gaze stability was measured in both axes separately and compared across conditions. Gaze was slightly more stable under high load in both axes: in the vertical axis, average deviation was 0.68° ($SD = 0.16^\circ$) under high load vs. 0.69° ($SD = 0.21$) under low load –

although this difference was not significant, $t(4) = 0.20$, $p = 0.84$. Similarly, in the horizontal axis gaze was steadier under high load ($M = 0.75^\circ$, $SD = 0.18^\circ$) than under low load ($M = 0.88^\circ$, $SD = 0.23^\circ$); again this difference was not significant, $t(4) = 1.75$, $p = 0.15$. The direction of this difference is the opposite of what would drive the hypothesised effect of perceptual load on orientation processing in any case, as with all else held equal, greater gaze stability would lead to more precise visual representations in distributed activity.

3.3.3.2 fMRI results

BOLD signal analysis. Activity of visually active voxels in early visual areas was compared between high and low perceptual load conditions. This analysis aims to confirm a general suppression of activity related to visual stimulation in early visual areas due to high perceptual load. A repeated-measures ANOVA was conducted on mean GLM parameter values extracted from the 100 most visually active voxels in each area, collapsed across presented orientations; load and visual area were within-subject factors in a 2 (low, high) by 3 (V1, V2, V3) factorial design.

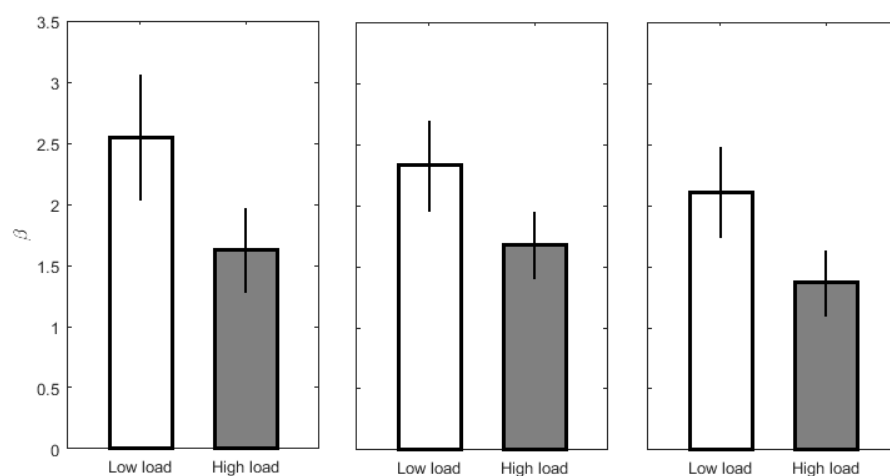


Figure 3-4. Mean GLM parameter values for visually responsive voxels in V1 (left), V2 (middle), and V3 (right) under low and high perceptual load conditions. Error bars represent \pm SEM across participants.

Analysis showed a main effect of perceptual load on BOLD activity, $F(1, 8) = 7.63$, $p < 0.05$; indicating that neural response to visual stimuli was suppressed across early visual cortex under conditions of high perceptual load, as can be seen in Figure 3-4, replicating the findings of Schwartz et al. (2005). The main effect of area was not significant, $F(2, 16) = 2.914$, $p > 0.05$; and the interaction between load and visual area was also not significant, $F(2, 16) = 2.94$, $p > 0.05$.

MVPA. A bank of 45 SVM classifiers were trained to discriminate a single orientation class from the rest; the classification accuracy of exemplars in the test set were calculated, averaged across cross-validation folds. Along with perceptual load, the other independent variable in our analysis was the absolute orientation offset between the true orientation of test set exemplars and the orientation prediction of the classifier. Therefore, to replicate the effects of Kamitani and Tong (2005) and demonstrate orientation selectivity in the voxels of early visual areas, it is expected that the classifier ensemble should be able to correctly predict the true orientation at levels above chance, while classification errors should occur most often for orientations adjacent to the actual orientation, with this decreasing as the offset increases. Identical analyses were carried for voxels extracted from V1, V2, and V3, under each load condition, independently.

In V1, there is was clear trend under both low and high perceptual load that positive classification of exemplars is reduced when the offset is increased (see Figure 3-5), demonstrating orientation selectivity across the set of voxels - distributed representations were more similar for similar orientations than for dissimilar orientations.

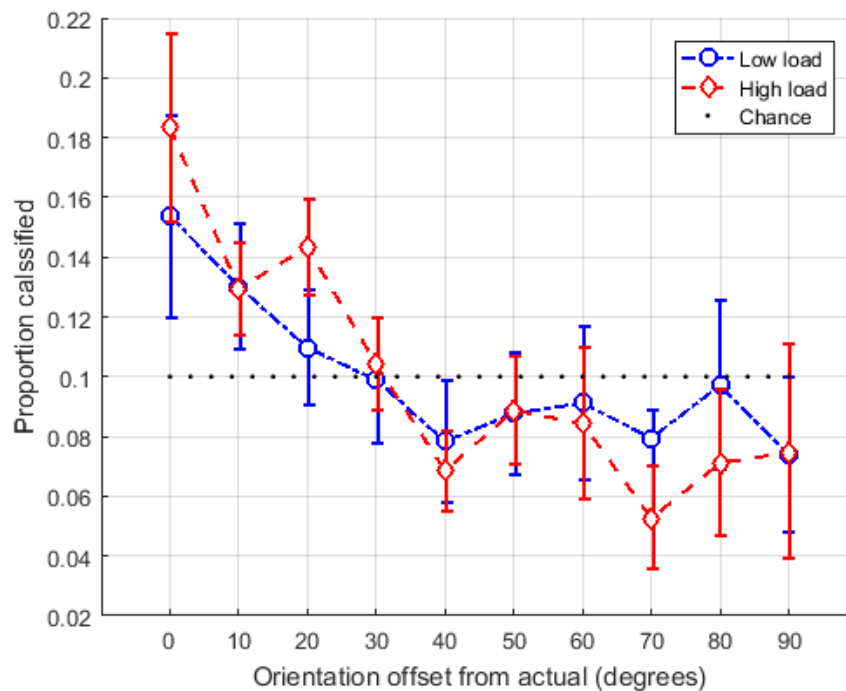


Figure 3-5. Classification performance for voxels in V1. A prediction with an orientation offset of 0° means the classifier predicted the correct orientation. Error bars represent \pm SEM across participants.

Note that accuracy is below chance towards 90° as the area under the curve must sum to 1; this arises as the classifier ensemble is only able to make one prediction per trial, so if the algorithm correctly predicts the correct orientation above chance then it must predict some other orientation with relatively lower probability. Replicating Kamitani and Tong (2005), it was found that the classifier was able to predict the correct orientation at levels above chance under both low and high load conditions by t-test: under low load, mean classification accuracy was 15.3%, significantly higher than chance (10%), $t(8) = 2.308$, $p < 0.05$; while under high load accuracy was 18.33%, again significantly higher than chance, $t(8) = 3.248$, $p < 0.05$. A paired t-test was conducted to compare classifier performance under low load vs. high load conditions, finding there was no significant difference, $t(8) = -0.782$, $p > 0.05$; the result does not support our prediction that perceptual load degrades the encoding of orientation in early visual cortex.

An alternative possibility however is that perceptual load interacts with orientation offset between the actual and predicted orientation class. While there may not be a reliable difference between load conditions in terms of actual orientation prediction, representations measured under high load may on average be misclassified as more distal orientations than under low load. As a specific example, a pattern induced by a 0° grating under high load could be misclassified as 40° , whereas the same grating may be misclassified as 10° under low load, which would indicate that distributed representations are more stable under low load. To investigate this, a two-way repeated-measures ANOVA was conducted across load and orientation. The load factor had 2 levels, while there were 10 levels of the orientation offset factor. By definition, the main effect of perceptual load was not significant, as the area under the classification accuracy curve necessarily sums to 1. A highly significant main effect of orientation was found however, $F(9, 72) = 7.25, p < 0.01$, confirming that distributed patterns of activity in V1 are tuned to orientation, as the classifier is more likely to make incorrect predictions which are close to the true orientation. The possible interaction between load and orientation was not significant $F(9, 72) = 0.68, p > .05$.

Identical analyses were conducted for visual areas V2 and V3 (see Figure 3-6). In V2, classification accuracy in both low load ($M = 14.86\%$) and high load ($M = 22.96\%$) was above chance, with $t(8) = 2.48, p < 0.05$, and $t(8) = 5.139, p < 0.001$, respectively.

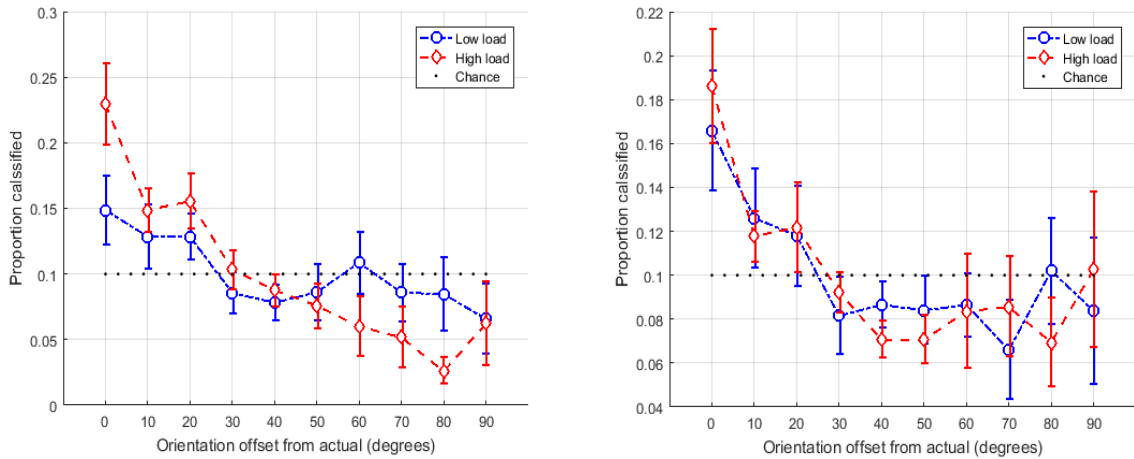


Figure 3-6. Classification results for voxels in V2 (left) and V3 (right).

Although accuracy was increased under high load, this difference was not significant, $t(8) = -2.018$, $p > 0.05$. Similarly in V3, classification accuracy was above chance: in low load ($M = 16.58\%$) with $t(8) = 3.165$, $p < 0.05$, and under high load ($M = 18.62\%$) with $t(8) = 4.049$, $p < 0.01$. Again there was no significant differences between low and high load classification accuracies by paired t-test, $t(8) = -0.792$, $p > 0.05$. Repeated measures analyses of variance across load and orientation prediction offset were conducted for classification results in both visual areas. The main effect of orientation was highly significant: in both V2, $F(9, 72) = 9.32$, $p < .001$; and V3, $F(9, 72) = 6.434$, $p < .001$, confirming the presence of orientation selectivity in these areas. Interactions between load and orientation were not significant however, $F(9, 72) = 1.79$, $p > 0.05$ in V2, and $F(9, 72) = .452$, $p = 0.05$ in V3.

3.4 Chapter Discussion

The aim of this chapter was to investigate whether deficits in the perception of orientation associated with high perceptual load can be traced to the degradation of distributed representations of orientation in early visual areas. Towards this end, in Experiment 1 a novel behavioural dual-task orientation

change detection experiment was designed. Consistent with the work of Stolte et al. (2014) it was found that under conditions of high load participants were less able to discriminate differently oriented gratings; confirming that the deficits associated with high load extend to fundamental visual features. Experiment 1 thus showed that perceptual deficits associated with load exist for the novel oriented grating stimuli used in the orientation change detection task, which were designed specifically with fMRI suitability in mind: the large radius and spatial frequency of the gratings being ideal for driving activity in visuo-cortical areas. In Experiment 2, the oriented grating stimuli of Experiment 1 were used in an fMRI experiment to measure the orientation-specific representational content of early visual areas. Several ancillary findings were replicated, namely the degradation of performance in the behavioural RSVP task under high load, reduced gross activity attributable to the peripheral grating stimuli in early visual areas under high load, and reliable orientation decoding in visual areas. However, our prediction that orientation-specific perceptual deficits under load can be traced to visuo-cortical representations was not confirmed, as MVPA methods found no reliable difference between difference between classification accuracy for oriented gratings presented under low or high load.

In Experiment 2, analyses confirmed that the central RSVP task successfully manipulated perceptual load: response accuracy was reduced and response latency was increased when subjects were required to respond to basic feature conjunctions in comparison to single feature targets. The effect of this manipulation was also confirmed on gross activity in early visual areas V1 to V3: activity induced by the experimental stimuli was reduced – replicating the general suppressive effect of perceptual load on visual activity as found in several previous studies (e.g. Rees et al., 1997; Schwartz et al., 2005). Furthermore, I used multivariate pattern analysis (MVPA) methods – namely a multiclass support-vector machine (SVM) classification ensemble – to reliably

predict at levels above chance the orientation of a presented grating stimulus using a set of 100 visually responsive voxels in each of the three visual areas investigated. This predictive ability was significant for each visual area, indicating that orientation information was encoded in distributed patterns of activity in each area independently. These findings are consistent with several previous studies reporting that fundamental visual characteristics can be inferred from brain activity (e.g. Kamitani & Tong, 2005; Haynes & Rees, 2006; Serences et al., 2009); furthermore this effect was present under conditions of high as well as low perceptual load.

However, further comparisons of MVPA efficacy between load conditions were not significant, yielding no support for the experimental hypothesis that fundamental feature representations in early visual cortex are modulated by perceptual load. This is especially surprising given the confirmation of perceptual load manipulation as implied by the reduction in overall stimulus-related neural response. There are perhaps aspects of the experimental design which could offer explanations of this null result – mechanisms which could act counter to the experimental hypothesis direction. One possibility is that the greater gaze stability found in the high load condition nullified a true perceptual load effect – although no statistically significant difference was found between gaze stability in low and high condition, this may be due to the relatively small sample size in the eye-tracking analysis. Another, more theoretical, possibility is that participants' spare perceptual capacity during central streams of low perceptual load was not allocated to *perceiving the orientation* of the surrounding grating. During such a state of passive viewing there is no guarantee that a person's perceptual priority list (as set and maintained by executive cognitive control systems under conditions of both low and high load; Lavie, 2000) places a certain stimulus or stimulus feature above others – in our case participants could have prioritised different aspects of the grating stimuli,

for example some apparent *motion* induced by the flickering of the grating stimuli rather than the orientation of the grating itself (e.g. Green, 1981; Sunny & von Muhlenen, 2014). This logic extends to a participant perhaps prioritising the perception of completely different stimuli altogether, separate from the grating itself, even within the mind's eye (Forster & Lavie, 2014). In each of these cases, the effect of prioritisation of non-orientation features is to reduce the information content of orientation representations in visual cortex – therefore if such mechanisms are active under conditions of low perceptual load then, even if high perceptual load does act to degrade orientation representations with all else held equal, there could be no measurable difference between representational content in visual cortex. Furthermore, the choice of 10 equally spaced orientations within the range $[0^\circ, 90^\circ]$ for the gratings restricted the analysis to MVPA; the related method of voxel tuning functions (VTFs; Serences and Saproo, 2009) may also have been informative regarding the modulation of orientation processing by load – however analysis of VTFs requires the presentation of gratings in the full 0° to 180° range of orientations. Therefore, in the next chapter, a new experimental design is employed which aims to account for these possibilities, and others related directly to physical stimulus parameters.

4 The effect of perceptual load on population coding of orientation

4.1 Chapter Introduction

Results of Experiment 1 in the previous chapter found behaviourally that the perception of orientation gratings was reduced under conditions of high load, however multivariate pattern analysis of brain activity in early visual areas found no statistical difference between orientation classification accuracy under low and high conditions. This inconsistency between MVPA and behavioural results is somewhat surprising given the work of Stolte et al. (2014), who psychophysically found that orientation perception was degraded by perceptual load. In this chapter I therefore set out to design a new fMRI experiment that incorporates behavioural measures of orientation perception. This addition is made to clarify whether the inconsistency between the previous MVPA and behavioural results can be attributed to characteristics of the previous experimental design which may have minimised the effect of load, or whether they indeed suggest that the origin of orientation-specific perceptual degradation under high load is not attributable to changes in low-level visuo-cortical representations.

One potential mechanism for the previous inconsistent results is the phenomenon of attentional capture. I hypothesised in the discussion of Chapter 3 that the flickering of grating stimuli could cause spare attentional resources in low load runs to spill-over to the perception of another feature of the display, such as apparent motion, rather than the grating orientation, nullifying any potential orientation-specific representational difference between the high and low load conditions. In addressing the possible capture of perceptual resources

by other facets of the stimulus, I introduce a dual-task paradigm, where a primary task modulates perceptual load while participants concurrently complete a secondary task which recruits orientation perception to leftover perceptual capacity – in our case participants must decide whether a presented grating is rotated clockwise or counter-clockwise relative to a sample grating presented previously. This task was employed previously by Kok et al. (2012) to investigate modulations of orientation representations due to prior expectation, and was shown to selectively recruit orientation processing in contrast to a task which only required stimulus contrast judgements. The addition of a secondary orientation-based task also allows direct monitoring of perception through behaviour; dual-task paradigms have previously been used to uncover the deleterious effect of perceptual load on detection of secondary-task stimuli (e.g. Carmel et al., 2007; Macdonald & Lavie, 2008). Therefore we will be able to confirm that the perception of behaviourally relevant visual information is indeed affected by the manipulation of load, whilst measuring the orientation-specific informational content of neural populations in early visual cortex. Given the hypothesised effect of load in degrading neural orientation tuning in line with the findings of Stolte et al. (2014), we expect high load to lead to an overall reduced ability to detect orientation changes. While there is also a possibility of an interaction effect such that perceptual load effects are more pronounced for smaller orientation offsets, detection of these small grating offsets may already be at floor level in low load, and hence be less sensitive to modulation by increasing the level of load.

In addition, in the previous chapter only on every other central cross presentation was a grating concurrently presented, meaning that for half of the grating presentations there was no cross task to perceive and exhaust perceptual capacity; any true effect perceptual load would therefore be more difficult to detect. In this chapter I therefore altered the procedure such that

oriented gratings were presented in full synchrony with the central cross task, and only for a brief period per orientation in an event-related design.

Furthermore, in this chapter I expand the set of oriented gratings to spread across the full range of possible orientations in equal steps; this allows the construction of voxel-based tuning functions (VTFs; Serences et al. 2009; Saproo & Serences., 2010) from cortical activity to characterise the orientation-specific tuning of neurons at the population level. Whilst operating on similar principles to MVPA regarding orientation encoding in the brain, namely leveraging orientation biases across neurons within individual voxels, these functions are more descriptive of tuning properties, allowing the extraction of specific factors such as orientation preference, response amplitude, and response bandwidth from estimated VTFs. These parameters form the basis of a more natural comparison between neural effects and the effects found psychophysically by Stolte et al. (2014) which concern orientation tuning specifically.

There are three potential ways in which Gaussian tuning curves may be altered by perceptual load - or indeed any experimental manipulation. The first is orientation-independent additive scaling, this describes a general modulation of neural activity in response to visual stimulation independent of the presented stimulus feature. This has been shown previously for perceptual load on task irrelevant stimuli in Experiment 2 of Chapter 3 and in previous work (e.g. Rees et al., 1997; Schwartz et al., 2005), where high load reduced BOLD signal related to task-irrelevant stimuli. As such it is expected that this form of scaling will be present for grating related activity here. However, a simple reduction in neural response signal, although certainly affecting our visual experience, may only be one part of the explanation. According to signal detection theory, for example, successful visual detection and discrimination do not only depend on the strength of the signal but also on the extent to which the signal is tuned to

the feature (Green & Swets, 1966). Therefore, through analysing properties of constructed VTFs under conditions of low and high load, I also investigate whether load induces orientation-dependent scaling of tuning curves, which can occur in two separate ways: multiplicative scaling and bandwidth scaling, see Figure 4-1 for depictions of both types.

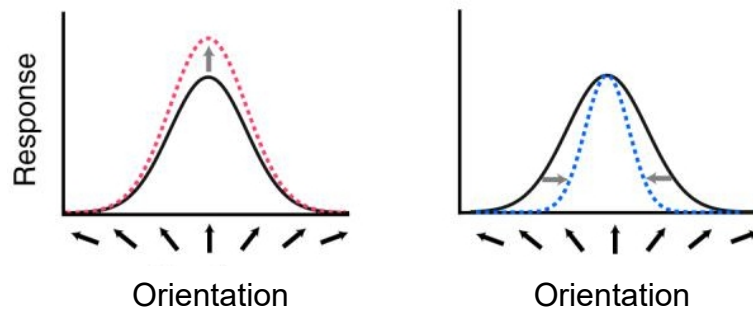


Figure 4-1. Multiplicative (left) and bandwidth (right) feature-dependent scaling of orientation tuning curves (adapted from Liu and Carrasco, 2009)

In Figure 4-1, the x-axis refers to orientation, and response profiles are depicted after centering at their preferred orientation. Multiplicative scaling refers to a linear change in response dependent on proximity of the presented orientation to the preferred orientation, which is equivalent to a change in tuning amplitude, while bandwidth scaling alters the selectivity of the population to the preferred orientation and is equivalent to a change in spread of the tuning curve.

4.2 Methods

Participants

Fourteen adults (mean age 22.9 years, standard deviation (SD) 3.5, range 19.3–35.6), 8 of which were male, participated in the study. All participants were recruited through the University College London subject pool, spoke English fluently, had no history of psychiatric or neurological disorders, and had normal

vision. Participants gave written informed consent and the study was approved by the University College London ethics committee

Experimental design

To investigate the effect of perceptual load on orientation tuning in early visual cortex a dual-task paradigm was employed: participants judged the offset of an oriented grating as clockwise (CW) or counter-clockwise (CCW) relative to a previously presented sample grating, while perceptual load was manipulated with a rapid serial visual presentation (RSVP) task at fixation. The central RSVP task was identical to that used in Chapter 3, however the inter-stimulus interval was reduced from 750ms to 500ms.

Concurrently with the central RSVP load task, participants completed a delayed orientation discrimination task in the periphery. During each trial a full contrast 'sample' oriented sinusoidal grating ($0.5 \text{ cycles/}^\circ$) subtending 10° of visual angle, with a 5° circular aperture removed to contain central task crosses, was presented 4 times in synchrony with cross presentation (250ms duration, 500ms interval, 4 repetitions). The orientation of the sample grating was drawn from 4 possible orientations evenly spaced across 180° (22.5° , 67.5° , 112.5° , and 157.5°) while the spatial phase of the grating was randomly selected from a set of ten possible phases on each presentation to attenuate the perception of apparent motion. Note that this orientation range imposes a limit on the minimum measurable bandwidth of voxel tuning functions of 45° at full-width-half-maximum, whereas a sampling from 8 evenly spaced orientation (as in Serences et al., 2009) would improve the resolution to 22.5° ; this decision was made to improve the signal to noise ratio of the measurements for each grating after confirming that the measured bandwidths in Serences et al. (2009) for VTFs in all visual areas exceeded 45° . The edges of the grating were smoothed

by convolution with a Gaussian kernel, so as not to elicit activity due to contrary orientation perception. Following the sample presentation was a 4.5s retention period where a 50% contrast Gaussian noise mask of equal dimension and position to the oriented grating was presented, also in synchrony with cross presentation (250ms duration, 500ms interval, 6 repetitions). A ‘test’ oriented grating of equal dimension, rotated CW or CCW relative to the previously presented sample grating (either $\pm 2^\circ$, $\pm 5^\circ$, $\pm 10^\circ$, or $\pm 20^\circ$), was then presented once for 250ms. There followed a number of noise mask presentations, synchronous to cross presentation and jittered between trials randomly from 4 to 7 repetitions, before the beginning of the next trial. A participant’s task on each trial was to indicate by a right-hand button press whether the test grating was rotated CW (key-press with ring-finger) or CCW (key-press with index-finger) relative to the sample grating; responses were recorded in a 3000ms temporal window following presentation of the test grating.

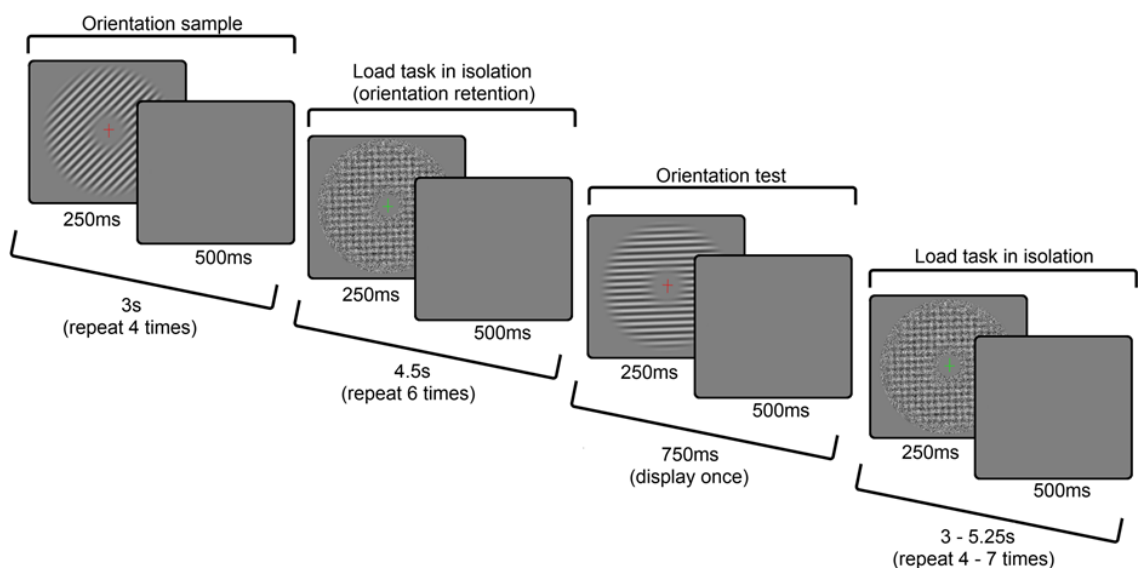


Figure 4-2. Schematic of an orientation discrimination trial during the experiment

Subjects completed 8 experimental scans, where each scan comprised a low-load RSVP stream and a high-load RSVP stream in succession, each stream beginning with a 3s instruction cue consisting of a fixation-dot at centre along with the 2 types of cross targets in the following stream (in low load: upright-red

and inverted-red crosses; in high-load: upright-yellow and inverted-green crosses), and ending with a 14s presentation of mid-gray with central fixation-dot to serve as a measure of baseline activity for region of interest selection. Across the whole experiment, each participant therefore completed 8 low-load and 8 high-load RSVP streams, alternating in an ABBABAAB fashion counterbalanced across participants. Within each stream, 16 orientation discrimination trials were run such that the 4 possible sample grating orientations were each presented 4 times; participants thus completed a total of 128 orientation discrimination trials under low perceptual load, and 128 trials under high load. Intertrial noise masks were presented for 4 to 7 repetitions randomly, under the constraint that each possible repetition amount occurred an equal number of times within each stream. The offsets of the test gratings were randomised across the whole 8 scans but were selected such that each offset had an equal number of overall occurrences in each load condition (e.g. the $+2^\circ$ offset occurred 16 times under low load and 16 times under high load).

Retinotopic mapping scans

For each participant, retinotopic mapping data was obtained with two separate functional scans using a combined ring and wedge stimulus. The procedure was identical to that described in Chapter 3.

fMRI and eye-tracking data acquisition

Eye-tracking and fMRI data acquisition equipment and parameters were identical to those specified in Experiment 2 of Chapter 3.

4.3 Data Analysis

fMRI image processing

All functional scans (including both retinotopic mapping scans) were preprocessed using the SPM8 software package in MATLAB (www.fil.ion.ucl.ac.uk/spm), using the same procedure as that described in Experiment 2, Chapter 3.

fMRI experiment modelling

Processed experimental scans were modelled as a generalised linear model (GLM) using the SPM8 software package. BOLD time-series for each scan for each participant were treated separately but underwent an identical GLM modelling procedure. Included in the model were event-related gamma function regressors for each sample grating orientation in each load condition. For example, in a single scan the 22.5° sample grating was presented under low load conditions 4 times (and flickered for 3s each presentation); this was modelled in the GLM as a regressor consisting of 4 gamma functions located temporally at the 4 grating stimulus onsets. Parameter estimates for these sample regressors after model estimation therefore represent activity elicited for each orientation under each load condition, resulting in a total of 8 sample orientation parameter volumes for each scan. Further to modelling the sample gratings, gamma function regressors were included to model the retention period (8 per scan, one for each load-condition X sample orientation combination), test grating presentation (2 per scan, one for each load condition), intertrial noise masks (2 per scan, one for each load condition), and instruction cues (1 per scan). Boxcar regressors were included to model the 14s mid-gray baseline presentation in each load condition (2 per scan), and the 6-dimensional continuous movement estimates extracted to realign EPI scans

were included as dummy regressors to account for potential movement artefacts. Therefore, in each scan 30 regressors were specified, which were then convolved with the canonical haemodynamic response function (HRF) and combined with a single intercept regressor to specify the GLM. This model specification was repeated for each scan and combined to form an overall session (i.e. 8 scans) GLM consisting of 240 regressors.

pRF mapping and ROI selection

The same population receptive field (pRF) estimation procedure as used in Experiment 2 in Chapter 3 was employed to retinotopically map the visual cortex; these maps were then used to isolate voxels belonging to V1, V2, and V3 visual areas. A similar region of interest (ROI) selection procedure was used, whereby the experimental scans served as a localiser. A GLM was specified with an event-related regressor (gamma function) marking the beginning of each sample oriented grating presentation and a regressor marking each 14s mid-gray baseline presentation. An SPM was calculated from a contrast of the sample grating regressor vs. the mid-gray regressor, and voxels sorted by t -value; the top 100 voxels by t -value within each visual area were then selected for subsequent analysis.

Multivariate pattern analysis

A one vs one multiclass classification ensemble method, identical to that used in Experiment 2 of Chapter 3, was employed to assess the representational content of distributed patterns in visual cortex. In this design however, there were 4 distinct sample orientations presented, resulting in an ensemble of 6 binary support vector machine (SVM) classifiers - the final classification of the system was again that orientation which elicited the most 'votes' from the

ensemble. Identical analysis was conducted independently for each retinotopically defined visual area (V1, V2, and V3). Input to the classifiers was the activity, as defined by the fitted GLM parameter values, of the 100 most visually active voxels within an area when presented with sample oriented grating stimuli. Sample orientation presentations (rather than test presentations) were used to avoid potential signal contamination with task anticipation or response movement. This resulted in a 100D vector representing the activity elicited by each orientation, for each load condition in each scan. Classifier performance was again collated over scans using a cross validation procedure, here with 8 scans in total: a classifier ensemble was learned using the representation vectors from 7 of the 8 scans, with performance measured as classifier accuracy on the 8th scan. Overall performance was given as the average classifier performance across the 8 permutations of the cross-validation procedure.

Voxel tuning functions (VTFs)

To construct VTFs (see Section 2.1.2.4 for details), data was prepared as parameter values of a GLM with each sample orientation block as a regressor, restricted to the 100 most visually active voxels, in an identical process to that described above for MVPA. Again, an identical analysis was conducted independently for each visual area (V1, V2, and V3). Each voxel within an ROI was assigned to an orientation preference based on the orientation which elicited the maximum response across 7 of the 8 scans, after the subtraction of mean responses to each orientation across voxels (to remove activations common to all orientations in each voxel). Using the remaining scan, the responses of voxels in each preference bin to each orientation were computed. This was repeated across all 8 cross-validation permutations for voxels within

each ROI. These data were then averaged to form preference-aligned orientation responses for each ROI.

4.4 Results

4.4.1 Behavioural results

RSVP task. For each participant, sensitivity indices (d') were extracted for high and low load conditions of the RSVP task according to

$$d' = Z(\text{correctdetectionrate}) - Z(\text{falsealarmrate})$$

where $Z(\cdot)$ refers to the inverse Gaussian cumulative distribution function. The mean d' in the high perceptual load condition ($M = 1.20$, $SD = 0.23$) was significantly lower than in the low-load condition ($M = 1.78$, $SD = 0.19$), $t(13) = 9.07$, $p < .001$. Reaction times for correct detections were also significantly longer in the high-load condition ($M = 664\text{ms}$, $SD = 61.89$) compared to the low-load condition ($M = 588\text{ms}$, $SD = 53.05$), $t(13) = -11.55$, $p < .001$. These findings confirm that the manipulation of load was effective.

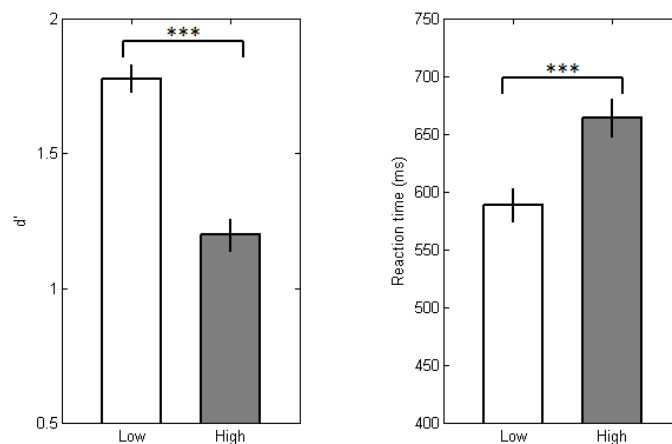


Figure 4-3. Sensitivity (left) and reaction time (right) for low and high load streams in the RSVP cross task. Both measures are significantly different between load conditions. Error bars represent \pm SEM.

Orientation discrimination. To confirm the degradation of orientation perception under high perceptual load at the behavioural level, offset direction (clockwise or counter-clockwise) discrimination accuracy was calculated for each offset magnitude (i.e. [2°, 5°, 10°, 20°]) in each load condition. Participants were given a 3000ms window following presentation of the test grating to respond to the task. If a response was not registered in this temporal window then that trial was not used for calculating overall accuracy; this occurred on only 53 trials across the entire experiment (1.4% of total trials). Mean discrimination accuracy across participants is shown in Figure 4-4.

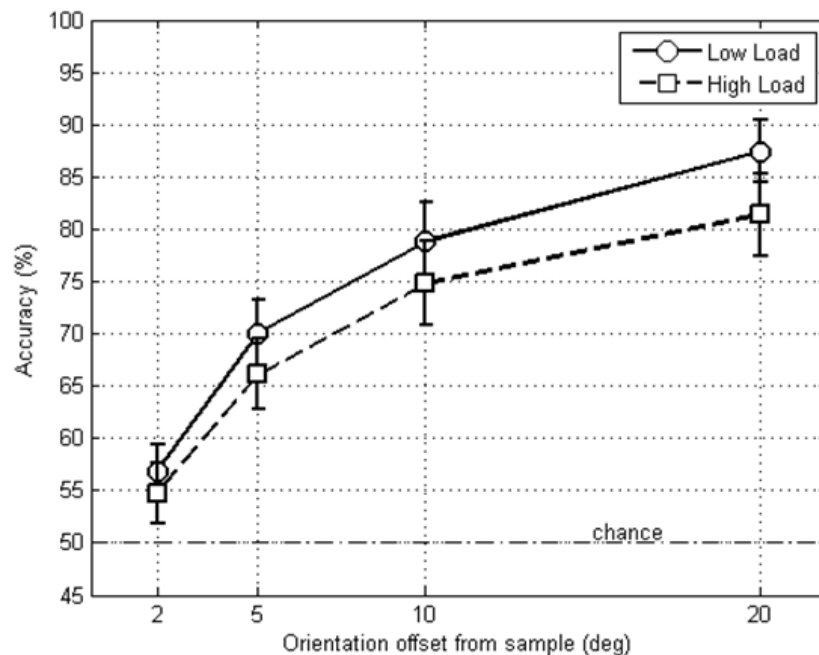


Figure 4-4. Orientation offset direction discrimination accuracies. Error bars indicate \pm SEM across participants.

A 4 X 2 (offset magnitude X load condition) repeated-measures ANOVA found a significant main effect of offset magnitude $F(3, 39) = 43.46, p < .001$, showing that perceptual judgements were harder for smaller orientation offsets, as expected. Importantly, the main effect of perceptual load was also significant, $F(1, 13) = 5.07, p < .05$, providing evidence that high foveal load inhibited

perceptual processing of the orientation stimulus. There was no interaction between offset magnitude and perceptual load, $F(3,39) = 0.36$, $p > .05$.

Eye-tracking. In order to exclude an alternative explanation of results in terms of gaze stability differences between load conditions. Due to technical difficulties, pupil position time-series for only 7 of the 14 participants were recorded and analysed. For the remaining subjects, eye position was monitored manually using the eye-tracker camera to ensure fixation. For the 7 participants we were able to collect eye-tracking data from, gaze was highly stable in both perceptual load conditions with a mean offset-from-fixation in low load of 0.35° ($SD = 0.16^\circ$), and a mean offset under high load 0.35° ($SD = 0.22^\circ$), the difference between conditions was not significant, $t(6) = 0.05$, $p = 0.96$. Gaze stability was measured in both axes separately and compared across conditions. Gaze was slightly more stable under high load in both axes: in the vertical axis, average deviation was 0.72° ($SD = 0.21^\circ$) under high load vs. 0.77° ($SD = 0.19$) under low load – although this difference was not significant, $t(6) = 1.8$, $p = 0.11$. In the horizontal axis gaze was significantly steadier under high load ($M = 0.61^\circ$, $SD = 0.13^\circ$) than under low load ($M = 0.79^\circ$, $SD = 0.19^\circ$): $t(6) = 2.80$, $p = 0.03$. The direction of this difference is the opposite of what would explain the modulatory effect on VTFs seen in V1 however, as increased variance in eye position would lead to broader tuning curves.

4.4.2 fMRI results

BOLD signal analysis. As in Experiment 2 of Chapter 3 here I compare the activity of visually active voxels across V1, V2, and V3 in response to visual stimulation.

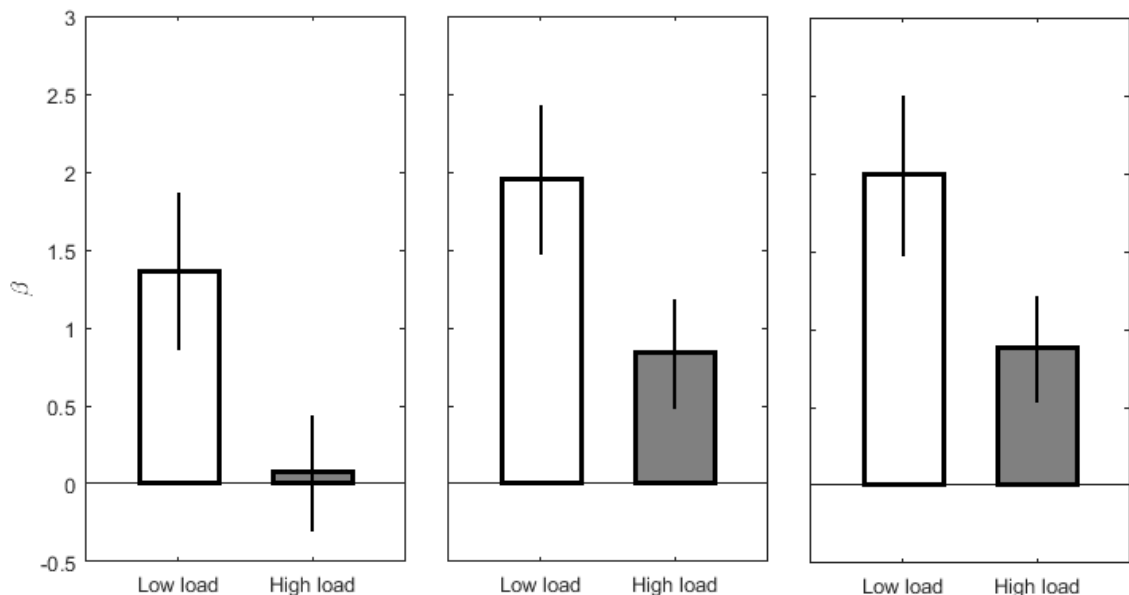


Figure 4-5. Mean GLM parameter values for visually responsive voxels in V1 (left), V2 (middle), and V3 (right) under low and high perceptual load conditions.

A 2 (perceptual load) by 3 (visual area) repeated-measures ANOVA showed a main effect of perceptual load on gross activity, $F(1, 8) = 6.753, p < 0.05$; again confirming that neural response to visual stimuli was suppressed across early visual cortex under conditions of high perceptual load. The main effect of area was not significant, $F(2, 16) = 3.182, p > 0.05$; and neither was the interaction between load and visual area, $F(2, 16) = 1.783, p > 0.05$.

Orientation tuning. Voxel-based tuning functions (VTFs) were constructed using GLM parameter estimates of the top 100 visually active voxels in each retinotopic visual area (V1, V2, and V3). VTFs were calculated separately for each load condition for each participant. An ANOVA of individual VTF values between load conditions was conducted to investigate the hypothesised feature-specific modulation of orientation response profiles due to perceptual load. A 4

X 2 (orientation offset X load condition) repeated measures ANOVA found a significant interaction between orientation and load, $F(3,39) = 3.08, p < .05$, for VTFs constructed using V1 responses, confirming a load induced modulation of orientation processing in early visual cortex. Identical analyses of VTFs constructed from V2 and V3 responses showed no significant interaction between grating orientation and perceptual load condition, $F(3,39) = 2.54, p > .05$; $F(3,39) = 1.78, p > .05$; respectively.

To investigate the nature of the load-induced orientation tuning modulation, population-wide VTFs for each load condition were characterised by fitting a circular Gaussian approximation (Von Mises function) of the form

$$f_{VM}(x) = \beta + \alpha e^{\kappa \cos(x-\mu)}$$

where β , α , κ , and μ correspond to baseline, amplitude, spread, and location related parameters, respectively. The Von Mises function was fitted to data collapsed across all participants (i.e. the grand average VTF) as robust fitting to every individual was not possible. The best fitting Von Mises function for the grand-average V1-derived VTFs across participants can be seen in Figure 4-6 below.

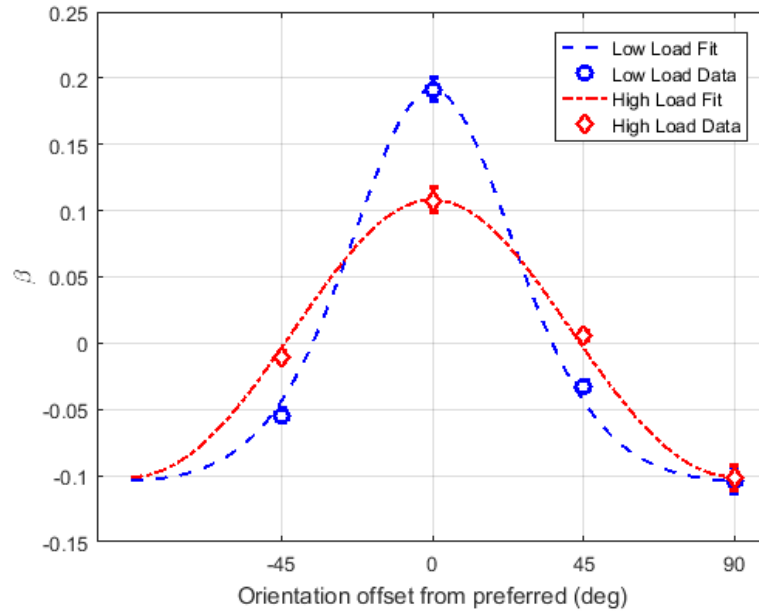


Figure 4-6. Population-wide VTFs (across 14 participants) in each load condition calculated using V1 voxels. VTFs are fitted with Von Mises functions. Error bars indicate \pm SEM across participants

The statistics of interest for the experimental hypotheses were the between-condition differences in response profile amplitude and spread, therefore response amplitudes were computed from population-wide VTFs as

$$A = \alpha(e^{\kappa} - e^{-\kappa})$$

and the spreads of VTFs in terms of full-width at half-maximum as

$$FWHM = 2\cos^{-1} \left[\frac{\ln \left(\frac{1}{2} e^{\kappa} + \frac{1}{2} e^{-\kappa} \right)}{\kappa} \right].$$

In V1, population-wide VTF response amplitude was reduced in the high perceptual load condition ($A_{high} = 0.20$) in comparison to the low load condition ($A_{low} = 0.29$), and FWHM was increased under high load ($FWHM_{high} = 86.08^{\circ}$) relative to low load ($FWHM_{low} = 57.36^{\circ}$). Both differences were confirmed as

statistically significant via nonparametric permutation test: condition labels were randomly permuted 100,000 times at the individual subject VTF level, with amplitude and spread differences, Δ_A and Δ_{FWHM} , being extracted from the resultant population-wide VTFs. The experimentally observed amplitude and spread differences were larger than 95% of permuted differences; $p = 0.044$ and $p = 0.040$, respectively (kernel density estimates of the null difference distributions can be seen in Figure 4-7), providing support for the experimental hypothesis that high perceptual load degrades orientation perception by increasing tuning width as well as reducing response amplitude in early visual

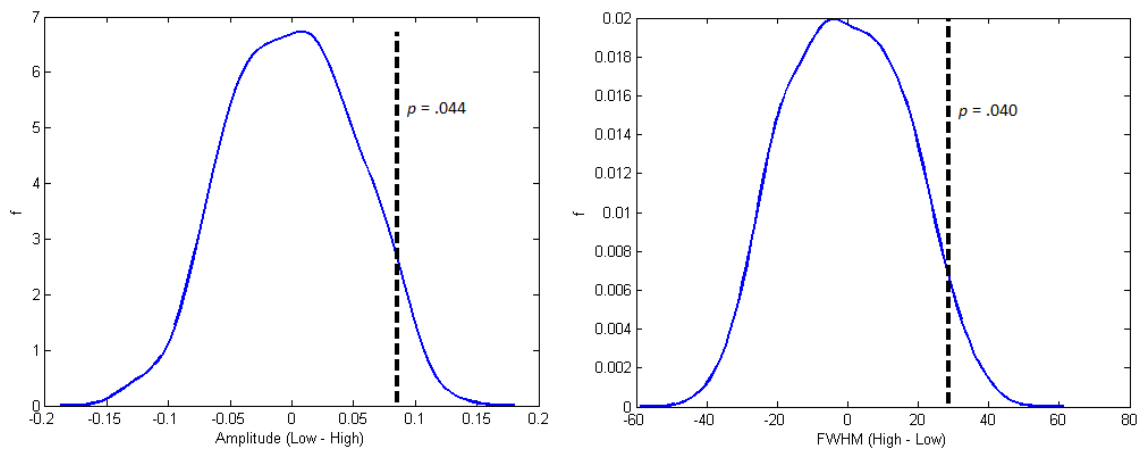


Figure 4-7. Gaussian kernel density estimates of the V1 VTF amplitude difference (left) and bandwidth difference (right) null distributions, calculated using 100,000 random condition label permutations. The black dashed line represents the experimentally observed values, and the associated p-value is reported.
cortex.

We also conducted an identical permutation test analysis for VTFs extracted from V2 and V3 activity (for fitted Von Mises tuning curves, see Figure 4-8). Consistent with the earlier ANOVA analysis on VTF values, VTFs extracted from V2 and V3 visual areas showed no modulation of either amplitude or spread parameters due to the perceptual load manipulation, both showing non-significant differences between conditions by permutation test, $p > 0.05$.

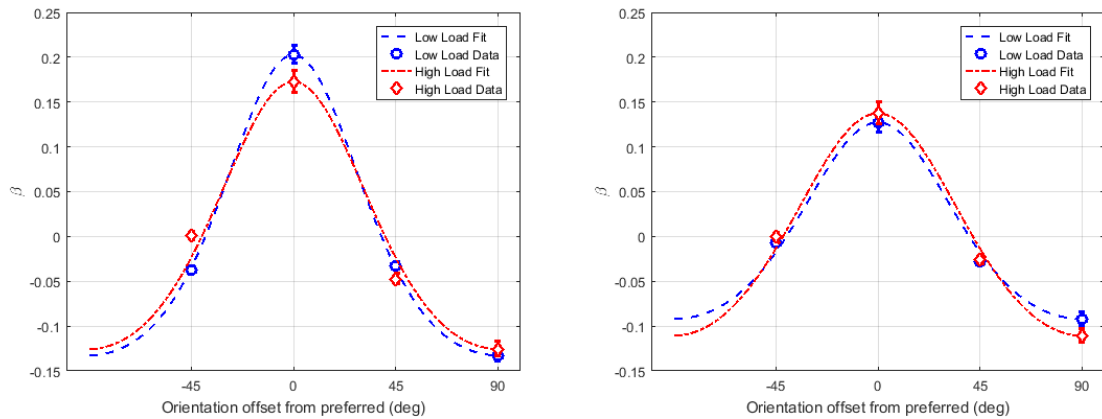


Figure 4-8. Population-wide VTFs (across 14 participants) in each load condition calculated using V2 (left) and V3 (right) voxels . VTFs are fitted with Von Mises functions. Error bars indicate \pm SEM across participants

Orientation preference distribution. The distribution of orientation preference for voxels in visual areas was calculated, after removing the mean signal at each voxel across orientations. Across visual areas, a higher proportion of voxels responded maximally to orientations near the horizontal axis (i.e. 22.5° and 157.5°) than those near the vertical. This is consistent with orientation preferences previously recorded in human V1 (using VTF analysis; Serences et al., 2009), mammal LGN and V1 (Sholl et al., 2013), as well as the *oblique effect* in human perception, where observers are more likely to perceive stimuli displayed at horizontal orientations rather than oblique orientations (e.g. Campbell et al., 1966, Furmanski and Engel, 2000; McMahon and MacLeod, 2003).

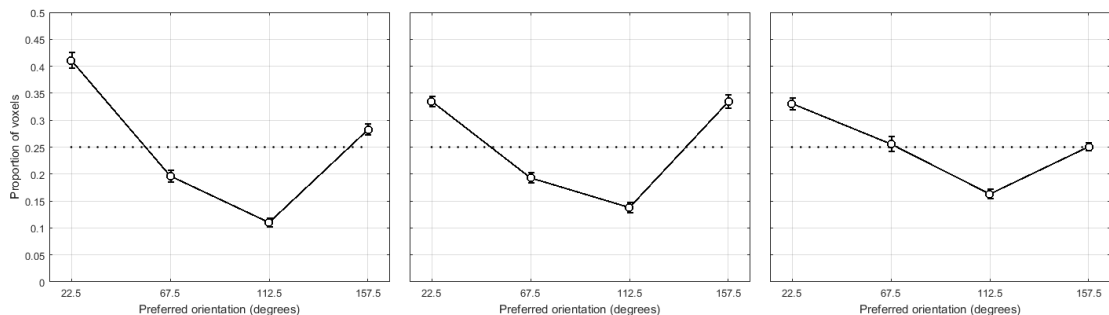


Figure 4-9. Distribution of orientation preferences in V1 (left), V2 (middle), and V3 (right). Error bars represent \pm SEM across participants, and the dotted line represents a uniform preference distribution across orientations

Orientation classification. An identical analysis to that conducted in Experiment 2 in Chapter 3 was carried out for visually active voxels across V1, V2, and V3. Figure 4-10 shows orientation classification performance for each visual area (note that offsets are not collapsed to *absolute* offsets, as in Chapter 3).

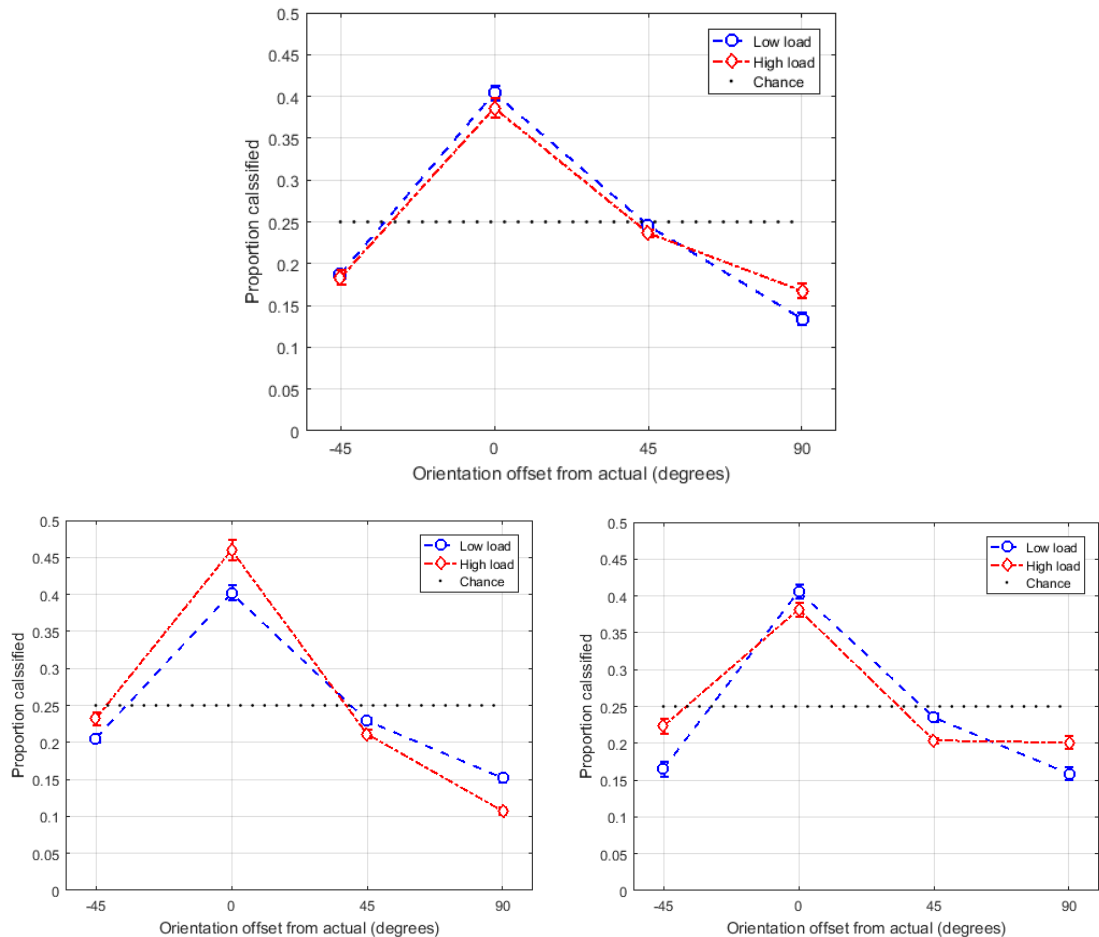


Figure 4-10. MVPA classification results for patterns extracted from V1 (top), V2 (bottom-left), and V3 (bottom-right) activity.

All visual areas displayed orientation selectivity, as evidenced by the rate of correct classifications being higher than chance under both low and high load conditions, and incorrect classifications being more likely to be assigned as an adjacent orientation (i.e. $\pm 45^\circ$ rather than $+90^\circ$). With regard to differences between load conditions however, while average prediction accuracy was reduced under high load conditions in V1, from 40.2% to 37.8%, this difference was not statistically significant, $t(13) = 0.915$, $p > 0.05$. Similarly, in V2 and V3

there was no significant difference in correct classification accuracy, $t(13) = -1.414$, $p > 0.05$, and $t(13) = 0.7054$, $p > 0.05$, respectively. This result, seemingly at odds with the VTF differences reported above, may be explained by the uneven distribution of orientation preferences for selected voxels in these visual areas – since there is a preponderance of voxels preferentially encoding near horizontal orientations, an increase in average individual voxel selectivity does not necessarily imply an increase in informational content for the area-wide representation.

4.5 Chapter Discussion

The overall findings of the chapter establish perceptual load as a unique factor in the modulation of perceptual and neural response to orientation. Perceptual load was successfully manipulated using a central RSVP task, as evidenced by reduced performance of the task in the high load condition, and by the gross suppression of grating-related visual activity in early visual cortex, consistent with previous fMRI results using the same or similar perceptual load modulations (e.g. Rees et al., 1997; Schwartz et al., 2005). Through the introduction of a secondary orientation discrimination task, behavioural measures of orientation perception were collected concurrently with fMRI data; while this task ensured spill-over of resources to the orientation feature of the task displays, in response to potential design criticisms of Experiment 2 of the previous chapter, analysis of these behavioural responses found reduced perception of orientation gratings under load, in agreement with work by Stolte et al (2014) and in Experiment 1 of Chapter 3 of this thesis.

Voxel tuning functions (VTFs; Serences et al., 2009; 2010) were constructed across the presented orientation range using BOLD responses recorded from

early visual cortex; independent VTFs were constructed using low and high load runs, and parameters of the fitted circular Gaussian tuning functions compared across conditions. VTFs constructed using V1 responses were shown to be of reduced amplitude and increased bandwidth under conditions of high perceptual load. This combination of multiplicative scaling (i.e. amplitude change), bandwidth scaling (i.e. width change), and additive scaling (i.e. gross suppression of activity) of population tuning curves represent a unique role for perceptual load in shaping feature-specific population neural responses. Spatial attention has been shown to multiplicatively and additively scale orientation tuning curves (e.g. Saproo & Serences., 2010; Kastner et al., 1999), and feature-based attention has been found to multiplicatively scale motion-direction tuning curves of single neurons and induce bandwidth scaling at the population level (Martinez-Trujillo & Treue, 2004), however no mechanism has previously reported affecting primitive feature tuning across all three possible types of curve scaling. VTFs constructed using V2 and V3 BOLD signals somewhat surprisingly showed no reliable difference between load conditions; this may be due to stimulus characteristics - e.g. the spatial frequency of the grating preferentially activating V1, as suggested by Kok et al., 2012 who found a similar pattern of results when manipulating stimulus expectancy - or the result may indeed reflect a true difference between the way higher-level attentional systems feed back to V1 in comparison to V2 and V3.

The reported modulation of V1 population response curves here is consistent with the effects found by Stolte et al. (2014) psychophysically. In their study participants detected a small vertical orientation grating against a background noise mask – the orientation offset of the noise mask relative to the grating and the contrast of the grating were varied systematically to produce behavioural tuning curves. Such curves were constructed under conditions of low and high primary task load; they found reduced orientation selectivity under high load,

indicated by increased tuning bandwidth and reduced amplitude. Our findings therefore suggest that the modulations of orientation perception by perceptual load found at the behavioural level can be traced, at least in part, to modulations at the earliest levels of cortical processing. The result is also consistent with that of de Haas and colleagues (2015) who found that the spatial selectivity of neural populations in early visual cortex were modulated similarly by perceptual load, where voxels responded to stimuli over a greater spatial extent under high load.

While the changes in experimental design implemented here in response to the findings and criticisms of Chapter 3 appear to have isolated the effect of perceptual load in cortical processing, a similar multivariate analysis (MVPA) of the results again showed no statistical difference between the pattern of activation across whole visual areas between low and high perceptual load conditions. This seemingly inconsistent finding may be explained through the methodological difference between VTF and MVPA techniques and the distribution of orientation preferences across voxels in the regions of interest. Using VTF analysis, it was found that voxels in V1, V2, and V3 responded preferentially to orientations near the horizontal meridian (i.e. 22.5° and 157.5°) in comparison to the vertical meridian, a result consistent with previous findings (e.g. Serences et al., 2009; Sholl et al., 2013). Therefore, whilst it may be true that tuning properties of individual voxels are on average modulated by load, this does not necessarily imply that the response pattern across a set of such voxels is reliably modulated also. As an extreme example, take the case where each voxel within the set is highly selective only for exactly 22.5° . Using VTF analysis it could be shown that the response of these voxels is on average slightly less selective to their preferred orientation under high load, however this tuning change may not be reflected in the distinctiveness between region-wide patterns induced by orientation gratings of 67.5° and 112.5° , for example, as

responses to both orientations would effectively be at baseline levels. Therefore, the unequal distribution of voxel preferences may render the MVPA method less sensitive to characterising orientation selectivity changes induced by perceptual load; a detriment which would not affect the VTF analysis which is based on the tuning properties of *individual* neurons. The VTF modulations in V1 presented here are therefore not contradicted by MVPA, however a systematic investigation into the effects of tuning curve parameters and orientation preference distributions on classification accuracy would shed more light on this hypothesis.

In conclusion, the results presented in this chapter establish a novel and unique role for perceptual load in shaping the perception of fundamental features of the visual scene. Orientation tuning curves constructed from neural responses in primary visual cortex were shown to be reduced in amplitude and precision, suggesting that previous perceptual deficits specific to orientation (e.g. Stolte et al., 2014), and indeed higher-level phenomena such as inattention blindness (Macdonald & Lavie, 2008), associated with increased perceptual load can be attributed to modulations of response profiles in neural populations at the earliest stages of visuo-cortical processing.

5 Modelling perceptual load in driving

5.1 Chapter Introduction

While it is well established that the perceptual load of a task determines the extent of attentional modulation of visual perception (Lavie, 2005; 2010; Lavie et al., 2015), the underlying factors of a given task which determine the level of perceptual load itself remain unclear. Experimental modulations of perceptual load rely on operational definitions set forth by Lavie (1995), who put forward task manipulations which are likely to load perceptual processes - for example increasing the number of items requiring memorisation in a given task. Therefore, manipulations such as the single-feature pop-out vs. feature-conjunction RSVP task used in previous chapters, target-distractor similarity, or visual-search set size have up to now formed the basis of investigations into the *effect* of varying perceptual load on perceptual processing, as the high-load conditions are expected to require greater perceptual resources in comparison to low-load conditions. A major motivation of the work presented here then, is to bridge this explanatory gap and characterise the relationship between the visual nature of a given task and the perceptual load induced by that task; that is, to shed light on the *cause* of variations in perceptual load.

A further motivation of the work is the austere nature of tasks hitherto used to modulate perceptual load. While manipulations involving stimulus features lend themselves well to numerical description and classification into low and high-load conditions, such displays, for example well-defined bars and unmoving letter configurations, are unlikely to form critical aspects of the rich dynamic tasks for which our perceptual systems evolved to complete. Therefore, in this work we aim to describe the modulation of perceptual load in real-world

dynamic scenes; not only does this improve the aforementioned ecological validity of the experimental stimuli and subsequent conclusions regarding the characteristics of perceptual load, due to the information-rich nature of the stimuli it should also allow a more gradual variation of perceptual load across stimuli, beyond discrete load categories used in previous investigations.

Perceptual load is not only a function of the visual scene however; the task being completed using the scene's visual information dictates the necessary amount of perceptual processing. In order to define the field of information over which perceptual load is determined then, a task must necessarily be defined: in previous empirical manipulations of load the tasks are defined explicitly along with the relevant units of the task themselves (e.g. the number of distractors in a visual search), and as such it is possible to vary the perceptual demands of the task directly. This precise approach is not feasible in the domain of real-world dynamic scenes however, therefore we opt to capture the observable visual scene during completion of a common dynamic task, specifically driving a car, from the point-of-view of the driver. This can then serve as information-rich and ecologically valid experimental stimuli in the context of a well-defined task recruiting perceptual processes.

Modelling the mental demands in the task of driving has up to now concentrated on the concept of *workload* which amalgamates several sources of load, making it impossible to disentangle the specific contribution of *perceptual* load.

Importantly for driving however, *perceptual* load is known to lead to inattentive blindness - and therefore a diminished ability to detect safety-critical events - while other types of workload may not have the same impact on driver perception, for example cognitive control load, that recruit higher-level executive functions such as working memory, in fact have the opposite effect in this regard (Lavie et al., 2004). The work here therefore aims to model directly the

relation between the visual scene during the task of driving and the perceptual load induced by that driving situation. A mathematical psychology approach is employed to this end (Coombs et al., 1970; Estes, 2014) where a purely mathematical relationship is derived between the visual information present during the task of driving and the psychological consequences, namely the level of perceptual load.

A machine-learning based methodology is used to discover this relationship between visual information and load. In essence a mapping function is learned using *labelled examples*; in the present work an example constitutes the pairing of a short video clip from the driver's point-of-view whilst driving in an urban environment with an associated numerical value representing the perceptual load of the captured driving situation. Using a corpus of such examples a regression analysis can uncover a mathematical relationship between videos and load values, through being trained to minimise estimation error across labelled examples. The learned model will therefore encode spatio-temporal visual information which is relevant for the prediction of perceptual load in the dynamic driving scene, such that when given a novel video clip of a driving situation, the model is able to estimate the perceptual load induced in the driver. The work therefore builds on that of Roper et al. (2013), who learned a mapping from performance measures and stimulus features in visual search to distractor effects due to perceptual load in a response competition paradigm, in two important ways. Firstly, we aim to recover perceptual load estimates directly from the visual scene itself (with the task of driving defined *a priori*), rather than through recourse to performance measures on other tasks, and secondly we extend the methodology to a visual task using complex dynamic information rather than the simple stimulus configurations explored by Roper et al. (2013).

In such an approach, it is necessary to compile a dataset of ground-truth video-label pairs for the mapping to be learned from. This process can be delineated into two sections: collecting relevant driving footage and labelling snippets of driving footage with a human-derived level of perceptual load. In the current work, driving situations are restricted to urban environments, and simple hypothesised task-units of driving, such as pedestrians and other vehicles, are varied across the videos - these factors serve as an initial basic criterion of perceptual load to guide video collection and to capture variance in theoretically relevant dimensions. Each collected video clip must then be assigned a ground truth label of perceptual load, and here we use the combined subjective judgements of a large number of subjects. The use of subjective labels are justified in this case as they are strongly correlated with objective physiological measures in several closely related domains (e.g. task-load, Mazur et al., 2013; driver workload, Marquart et al., 2015). After the collection of this ground-truth dataset, a variety of state-of-the-art computer vision and machine-learning techniques are investigated and developed in a competitive analysis in describing the relationship between dynamic visual driving scenes and the associated perceptual load most accurately.

5.2 Methods

5.2.1 Building a video dataset

Data recording equipment

The data collection vehicle was a Toyota Prius equipped with a high-quality dashboard mounted camera (Point Grey Flea3 model) and high-precision global positioning system (GPS). The camera was centrally placed on the dashboard facing forwards and captured 75° of visual angle at 30 frames per second. No

zooming, focus, or gain adjustments were made during recording, focus was set at infinity, and the gain and shutter speed were locked. Camera aperture was opened at the beginning of each recording session as much as possible without allowing white objects in the scene to saturate. The recorded raw high-resolution images were later compressed using *ffmpeg* to a MPEG 4 Part 14 video format at a resolution of 640 x 512 pixels. The GPS device recorded time-stamped longitude, latitude, and altitude data at an average precision of .5m at a rate of 180Hz, synchronised with the camera shutter (6 GPS samples per video frame).



Figure 5-1. Example frames from captured video in Brussels city centre.

Data collection routes

Two data collection routes were designed in and around central Brussels, Belgium. Routes were designed to capture variation in vehicle and pedestrian density throughout the day, and contained a variety of common urban road types: intersections, junctions, roundabouts, and straight roads.

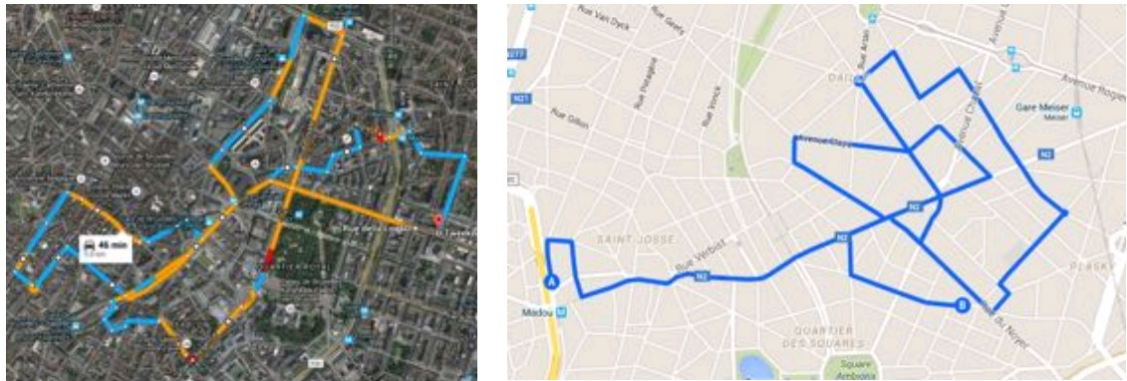


Figure 5-2. Google maps screenshots of the 2 planned routes in central Brussels. Route 1 (left) took an average of 60 minutes to complete, while route 2 (right) took an average 40 min.

Each route was completed 5 times on separate, fine-weather, days. The table below reports the beginning and end time-stamps for each data collection run on each route. The 10 total runs resulted in the collection of over 12 hours of high-quality video and GPS data.

Sequence labelling

Each collected video was then viewed and manually partitioned into individual sequences according to several features of the driving situation. Any periods of very slow ego-motion were removed from the dataset (i.e. the data collection car travelling at a speed of less than approximately 5 miles per hour). There were 6 features used to describe the videos, which are detailed in Table 5-1.

Table 5-1. Features of the driving scene used to describe captured video and partition into individual sequences.

<i>Feature</i>	<i>Possible values</i>
Current road layout	Straight road; intersection (including junctions); roundabout
Carriageway type	Dual or single
Number of lanes	Integer value (from 1)
Current ego-car manoeuvre	None; right turn; left turn
Pedestrian density	Integer value from 0 (no pedestrians in view) to 3 (large numbers of pedestrians)
Vehicle density	Integer value from 0 (no vehicles in view) to 3 (large numbers of vehicles)

A new sequence was declared and labelled when one or more of the features of the scene changed from the previous sequence. For example, if a group of pedestrians appeared on the pavement after exiting a building, where previously there had been no pedestrians in view, then, all else in the scene being equal, a new sequence was declared and the pedestrian density value increased from 0 to a higher value (depending on the number of pedestrians). Through this system number of sequences were created, each labelled with the 6 features described above. The length of the partitioned sequences ranged from 2s to 18s.

Clip selection

Given the labelled sequences, a heuristic method was implemented to further partition the sequences into a selection of 2s video clips which would become the experimental dataset. Two-second clips of a sequence were more likely to

be included in the dataset if they formed a grouping with clips from other sequences recorded at the same location; 2s clip groups were then more likely to be included if there was a high variance of pedestrian and/or vehicle density within that group of clips. Groupings of 2s clips at a single location were formed using GPS data: if the ego-car position was within 10m for a duration of at least 1s across a pair (or more) of sequences then 2s clips were extracted from those sequences and formed a group at that location. Each group was then given a score dependent on the variance of pedestrian and vehicle densities of clips within that group:

$$score(G) = G \cdot (var_G[d_p] + var_G[d_v]),$$

where G refers to the clip group, and d_p and d_v refer to pedestrian and vehicle densities, respectively. The final dataset was then selected as the set of clips which maximised this score across possible groupings in a greedy fashion, resulting in a total of 1809 distinct 2s clips. Figure 5-3 displays descriptive statistics of the data set; the number of videos per location type and the number of location matched videos per size of location group.

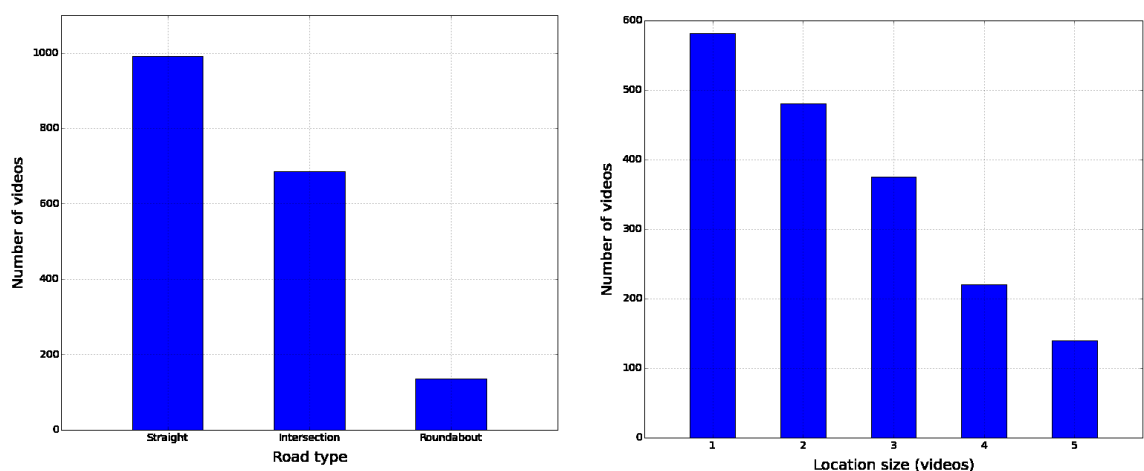


Figure 5-3. On the left, the types of road situation in the dataset by frequency, and on the right the number of videos per location group size (e.g. there were 488 videos matched with one other video at the same location).

5.2.2 Estimating ground-truth values of perceptual load

With the aim of assigning perceptual load values to each of the collected video clips, a pairwise comparison method was used. In this paradigm (e.g. Thurstone, 1927; Bradley & Terry, 1952; Luce, 1959), participants are presented with a pair of stimuli and are prompted to indicate which stimulus has the greater amount of some subjective attribute (in our case, perceptual load). After the collection of many such pairwise judgements, a value along the attribute of interest can be assigned to each stimulus by fitting a probabilistic model of comparison outcomes; common models being the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959), Elo (as used to rate chess players; Elo, 1978), and TrueSkill (Herbrich et al., 2006). The TrueSkill algorithm (described in Section 2.2.1) is used here as it has shown state-of-the art performance in similar domains (Chen et al., 2013).

Participants and experimental platform

For accuracy in the method each video clip was necessarily compared many times. To enable this, the comparisons were crowdsourced: 83 participants were recruited via crowdsourcing company Pallas Ludens and paid 20EUR/hour for participation. Each participant performed pairwise comparisons for 2 1-hour sessions on separate days and performed the comparison tasks under the supervision of Pallas Ludens at a facility in Germany. It was ensured that participants held a full driver's license. A web-application was written to deliver the comparison task interface to the remote participants through a web-browser. Participants were sat at IBM PCs, with 24" monitors, equipped with Google Chrome software to view the pairwise comparison web-application, and they were instructed to remain roughly 0.5m from the screen while viewing the video clips.

Experimental design

Subjects viewed pairs of video clips and were instructed to indicate which situation depicted by the video clips would require the greatest demand on attention if they were driving in that situation. This concept of attentional demand fits the operational definition of perceptual load put forth by Lavie (1995) and is readily explained to laymen. In verbal instructions to the participants this was also explained by example, for instance: “in which driving situation would you be more likely to hush a talking passenger?”

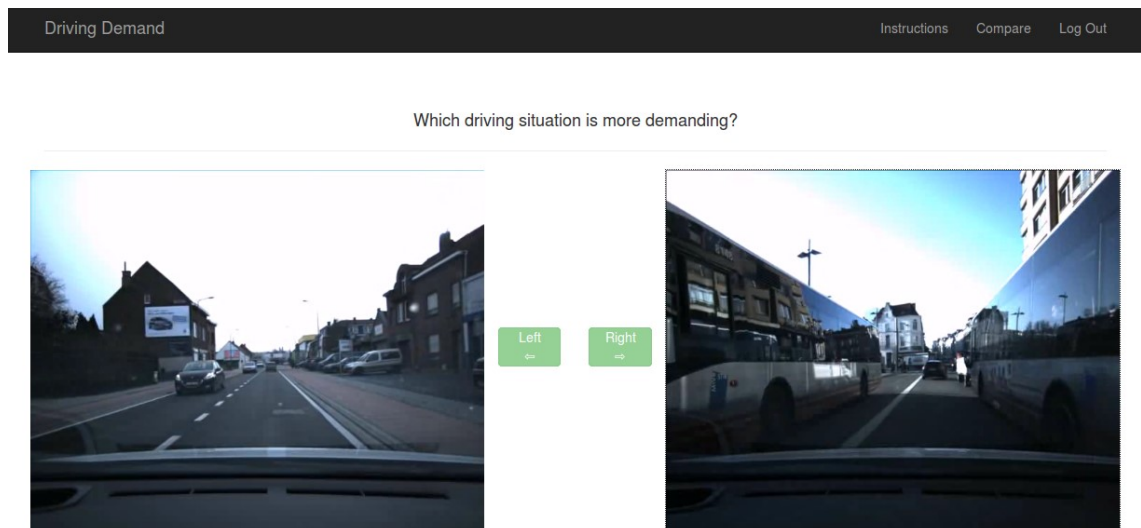


Figure 5-4. Layout of experimental interface.

The experimental interface was a web-application written in the Python programming language (using the Flask package), and was served to PCs through Google Chrome. For each comparison, two driving clips were presented next to each other horizontally. Each video was presented with a width of 9cm across all PCs and participants. The clips did not play automatically, rather the participant was required to manually press play on the video player for each video; the videos could also not be played concurrently to ensure participants were attending to only one video at a time. Participants

could not select the video they deemed more demanding until each video had been viewed at least once; this response was indicated by pressing a button corresponding to the 'left' or 'right' video on the screen. Videos were required to be played in alternating fashion, and the maximal number of views was 5 - at which point a response was forced.

Selection of the video pairs presented to the participants was randomised in 35 rounds. Each round contained 1809 comparisons (i.e. each video being involved in 2 comparisons per round), and comparison tasks were allocated to participants on a first-come-first-serve basis - when a participant completed a comparison of a video pair, they were then served with the next video pair to compare immediately. Randomisation was achieved by first representing the round of comparisons as an 1809 X 1809 binary matrix: a 1 at position (i, j) in the matrix represented a comparison between the i th and j th video clip in the data set. This comparison matrix was initialised as a diagonal matrix; the final randomised matrix was then realised by randomly permuting the rows of the initial diagonal matrix. This randomisation procedure was carried out at the beginning of each round before the comparisons were placed in a queue to serve to participants. This resulted in each video being compared to another 70 times, resulting in a total of 63,315 $(1809 * 70/2)$ driving situation comparisons.

5.3 Results 1: Ground-truth perceptual load values

Due to a malfunction of the Pallas Ludens video server, the first 3 rounds of comparisons (i.e. 5,427 comparisons in total) were removed from the data set before estimation of perceptual load values. These initial comparisons took on average 45s for participants to complete as there was a large lag in video loading - we therefore exclude them due to the potential case where a

participant would view the video in short snippets as it loaded and therefore would not gain a full sense of the driving situation. This removal resulted in a total of 57,888 video comparisons.

Distribution of load in the dataset. The TrueSkill algorithm was applied to the collected pairwise comparisons to arrive at an estimate of perceptual load level for each video depiction of a driving situation. Each video's load value was initialised at 25 with a standard deviation of 8. Figure 5-5 shows a histogram of perceptual load values after all comparisons were processed by the algorithm; it is bell-shaped, indicating that perceptual load on the road is distributed in a Gaussian fashion.

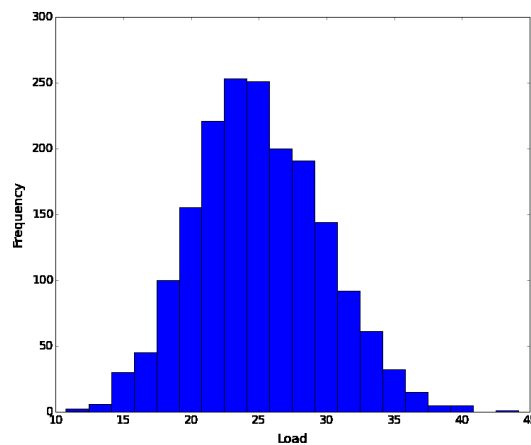


Figure 5-5. Histogram of perceptual load values as estimated by the TrueSkill algorithm.

Rating stability. A correlation analysis was carried out to investigate whether the final load values obtained were reflective of the true distribution in the dataset, and thus confirm that the number of comparisons was sufficient. In this analysis the correlation coefficient, $Corr(x, y)$, was computed between a vector of the concatenated load values at comparison round n , W_n , and a vector of concatenated load values at the previous comparison round, W_{n-1} , such that $Corr(W_n, W_{n-1})$ quantifies the similarity between the estimated load values on successive comparison rounds. If sufficient comparison rounds were included,

then this value should approach 1 asymptotically, and should very nearly reach 1 during the last comparison rounds.

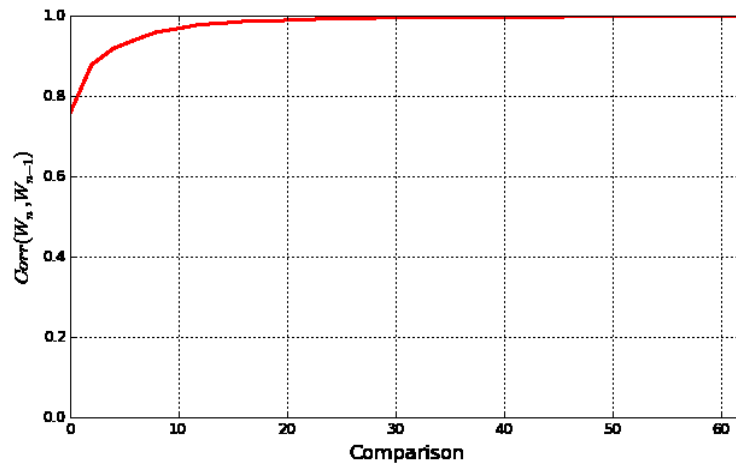


Figure 5-6. Correlation of the perceptual load values across the whole dataset between comparison rounds

As can be seen in Figure 5-6, this is indeed the case, with a correlation coefficient of 0.99 between the perceptual load values at the final round and penultimate round. This result indicates strongly that the perceptual load values as estimated by the TrueSkill algorithm converged upon the most likely distribution of perceptual load.

In summary then, we have compiled a dataset of 1809 video clips, each being associated with an estimate of perceptual load in the scene. The next section sets forth the method used to predict the perceptual load value given a video - this consists of a video feature extraction stage and a regression stage.

5.4 Data analysis

5.4.1 Video representation

Each video clip was processed to extract information-dense spatio-temporal features, using both the improved dense trajectory (IDT; Wang et al., 2013,

2015) and convolutional 3D (C3D; Du Tran et al., 2014) pipelines. The descriptor extraction and fusion procedures are described in detail in Section 2.2.2.

In the current formulation of IDT, parameters were chosen as having previously produced the best performance in human action recognition experiments (Wang et al., 2013). Each of the videos in the dataset was therefore described by 5 bag-of-visual-word histograms, each of 4000 dimensions. The 5 histograms vectors correspond to: trajectory displacements, histogram of oriented gradients (HOG; Dalal & Triggs, 2004), histogram of optical flow (HOF; Wang et al., 2011), and motion boundary histograms in each spatial axis (MBHx and MBHy; Wang et al., 2011). To extract the IDT descriptors, C++ source code made available alongside the original IDT publication (Wang et al., 2013) was used with a sampling stride of 5 pixels. Subsequent descriptor normalisation and the bag-of-visual-words pipeline (both learning and assignment) were implemented using the Python language (equipped with the ‘scientific stack’ of packages: numpy, scikit-learn, and Theano for GPU computation).



Figure 5-7. Example dense trajectories in our dataset. The red points indicate sampled trajectory positions in the current frame, and the green trails indicate their locations in previous frames.

C3D features (Du Tran et al., 2014) were extracted using a pre-trained 7-layer 3-dimensional convolutional neural network. The network architecture consisted of 5 combination convolution-max-pooling layers followed by two densely

connected layers: the descriptor we extract for each given video is taken from the last fully-connected network layer (layer 7), which is a 4096D vector. The network was realised using the Caffe deep learning framework with network weights pretrained on the Sports1M dataset (obtained from the Dartmouth College repository; see Bibliography under Sports1MDownload).

For each video then, 5 IDT feature vectors, each being 4000D, and a C3D feature vector of 4096D, were extracted. In the original application of IDT to action recognition (Wang et al., 2013), the 5 channels of information were combined using a χ^2 multichannel kernel for input to an SVM classifier, while in the C3D formulation (Du Tran et al., 2014) the single 4096D feature vector was used as input to an SVM classifier. In this work we compare these configurations to a novel IDT+C3D representation, which aggregates the high motion information content of IDT, which pools trajectory-guided features across 15 frames of video, with C3D features which are biased towards static appearance information, given that temporal information is only pooled across 3 frames in each convolution. Through framing the C3D descriptor as an additional feature channel, the IDT and C3D descriptors are combined with a multichannel kernel, with a χ^2 kernel being used for each individual IDT channel, and radial-basis function (RBF) kernel being used for the C3D channel. The individual kernel matrices are then normalised and averaged. This nonlinear kernel fusion was also compared against a baseline of linear kernel fusion, where each IDT and C3D channel was subject to a linear kernel. In each case the representation of a video after fusion is a vector of numerical similarities to all other videos in the dataset used to construct the kernel matrix. For example, in the case of training on the whole 1206 exemplar training set, each video is represented as a 1206D vector.

5.4.2 Regression and model fitting

Multichannel kernel and model configurations

The regression analysis consists of a competitive comparison of IDT, C3D, and IDT+C3D feature fusion and regression pipelines. Two broad regression models are investigated for IDT+C3D representations: support vector regression (SVR) and ridge regression (see Section 2.2.3 for details), along with two types of multichannel kernel functions for channel fusion: linear per channel and the χ^2 + RBF nonlinear configuration described above. This results in 4 initial model configurations whose results are reported.

Further analysis of the IDT+C3D descriptor explores the potential of weighting the individual channel kernel matrices, such that the multichannel kernel equation becomes

$$K(x_i, x_j) = \sum_c W^c \cdot k^c(x_i^c, x_j^c) / A^c$$

where W^c is the weight of the c th channel and is a tunable parameter of the system. This enables the investigation of whether certain feature types are more informative for predicting perceptual load as well as if this increased model freedom produces a more accurate overall load prediction system.

Hyperparameter optimisation

For each feature channel fusion and regression pipeline there exist hyperparameters which affect its performance, for example the λ parameter of a ridge regression dictates the extent of regularisation in the model. To tune these

parameters we use a tree of Parzen estimators (TPE; Bergstra et al., 2011) sequential model based optimisation (SMBO) procedure (see Section 2.2.4 for details). For each of the feature, fusion, and regression configurations there exist hyperparameters which are tuned using SMBO, which can be seen in Table 5-2.

Table 5-2. Tunable hyperparameters of the pipeline configurations in the current experiments.

Features	Kernel	Regression	Tunable hyperparameters
IDT	χ^2 nonlinear kernel	SVR	6: C (regression penalty) and γ (kernel width) for each channel
C3D	RBF nonlinear kernel	SVR	2: C and γ
IDT+C3D	Linear kernel	SVR	1: C
IDT+C3D	Linear kernel	Ridge	1: λ
IDT+C3D	χ^2 + RBF nonlinear kernel	SVR	7: C and γ for each channel
IDT+C3D	χ^2 + RBF nonlinear kernel	Ridge	7: λ and γ for each channel
IDT+C3D	χ^2 + RBF nonlinear kernel - fixed kernel width with tunable channel weights	Ridge	7: λ and W (channel weights) for each channel

For each system configuration in Table 5-2, the initial sampling distribution for each tuneable hyperparameter was set as the lognormal distribution with a mean of 0 and standard deviation of 1. Figure 5-8 portrays the lognormal distribution parameterised in various ways. An exception is the configuration with tuneable channel weights, which are instead sampled from a uniform distribution with minimum 0 and maximum 1. As the sequential TPE-based

SMBO progresses, the initial sampling distribution for each parameter is adjusted so as to favour values more likely to maximise expected improvement (EI; Jones, 2001).

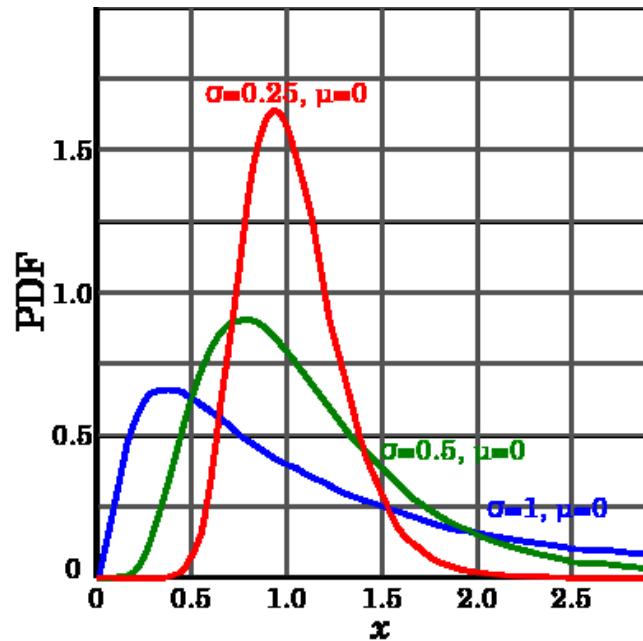


Figure 5-8. The lognormal distribution, parameterised by a mean, μ , and standard deviation, σ , which describes a continuous random variable whose logarithm is normally distributed. The configuration used in throughout is in blue ($\mu=0$, $\sigma=1$).

In the current experiments, the SMBO algorithm maximised R^2 , i.e. the coefficient of determination of the fitted model, or the amount of variation in the true perceptual load values which is explained by the model. The value of R^2 at each SMBO iteration was calculated using a 3-fold cross validation procedure on the 1206 example training set: 2/3rds of the data (i.e. 804 exemplars) was used to train the model, and model predictions were generated for the remaining 1/3rd of examples, R^2 was then computed between the model predictions and the ground-truth perceptual load values of those examples. This value was computed for each of the three permutations of cross-validation training and test splits and averaged across splits to give the SMBO model performance for that hyperparameter configuration.

5.5 Results 2: Predicting load

In the following sections, regression results on our driving video data set are presented for each descriptor-fusion-regression combination shown in Table 5-2. The performance of the original IDT and C3D configurations is reported first, followed by the performance of the combined IDT+C3D descriptor in combination with the several kernel and regression types.

5.5.1 Original IDT and C3D configurations

Original C3D method. Here raw C3D descriptors are used in combination with support vector regression. The SVR error penalty parameter, C , was optimised using the TPE-based SMBO algorithm described in Section 2.2.4. SMBO was run for 500 iterations. The progress of the SMBO algorithm in terms of best model performance over the 500 iterations is shown below.

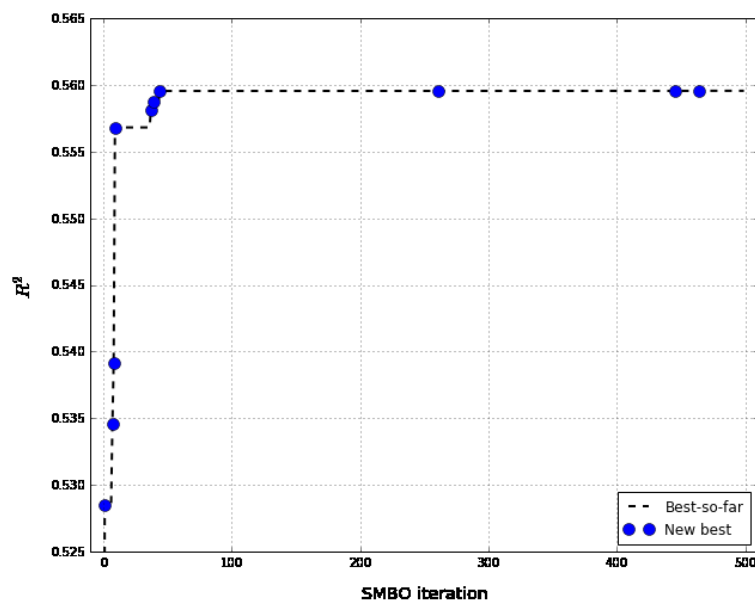


Figure 5-9. Progress of the SMBO algorithm in terms of model variance explained across the 500 algorithm iterations. Blue dot markers represent a new best configuration being discovered by the algorithm

The best SVR penalty parameter was discovered on the 453rd SMBO iteration and corresponded to $C = 3.1$, with an average cross-validation R^2 of .558. This value is the average performance across the 3 cross-validation folds used in SMBO, therefore to estimate the configuration's generalisation performance, it was trained using the full (1206 example) dataset. Perceptual load predictions were then made on the held out validation set (603 examples); R^2 was then calculated between the predictions and the true perceptual load values. A model prediction vs. ground truth load value plot, displaying the true and predicted load value for each test set exemplar, is given in Figure 5-10

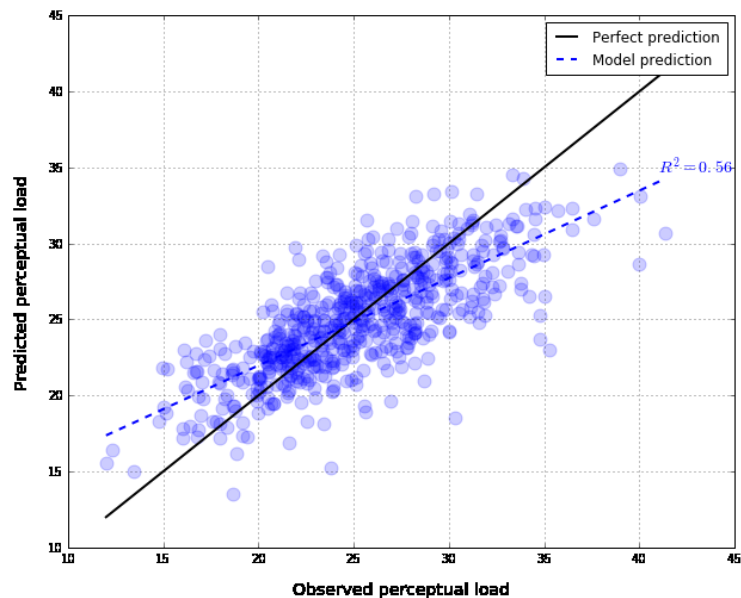


Figure 5-10. Each blue marker represents a test set exemplar - its position on the x-axis is the ground-truth TrueSkill estimate of perceptual load, while the y-axis position is its predicted perceptual load according to an SVR model using raw C3D features, trained on the full training set (1206 examples). The $y = x$ black line represents a model with perfect predictive power (i.e. 100% of variance explained by the model); the dotted blue line represents the fit of the trained ridge regression model

SVR with $C = 3.1$ trained on the full training set accounts for 56% of the variation in perceptual load values of the held out validation set ($R^2 = 0.56$). As can be seen in Figure 5-10, the model accounts for the general trend in the data, however it is biased to predict towards the mean of the data, i.e. extremely low load examples are estimated to have a higher level of perceptual load, and the same phenomenon is visible for extremely high load examples.

Original IDT method. Here the performance of the 5-channel IDT descriptors in combination SVR is reported. During 500 iterations of SMBO, the SVR error penalty parameter, C , along with the individual channel χ^2 kernel widths (one for each IDT feature channel), were optimised.

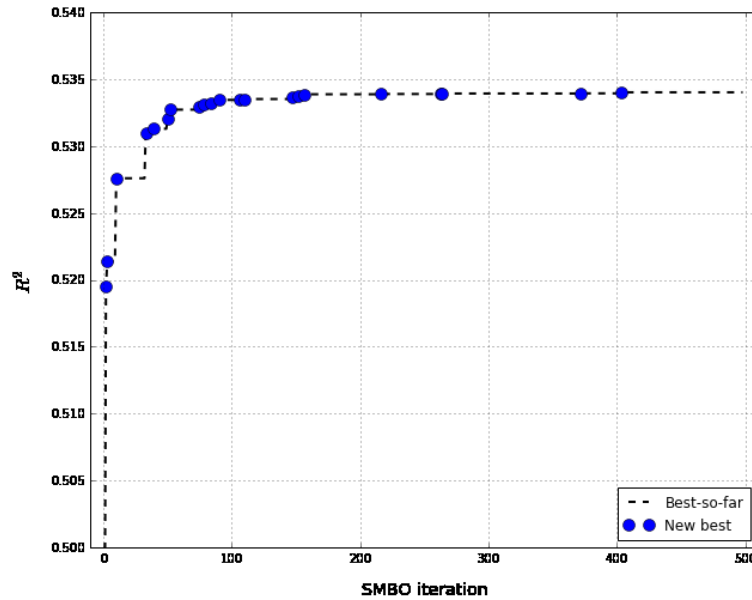


Figure 5-11. Progress of SMBO for SVR with nonlinear χ^2 kernel fusion of IDT channels

The best performing IDT configuration was realised on the 403rd iteration of the SMBO algorithm, with a cross-validation R^2 of 0.534 (see Table 5-3).

Table 5-3. Best configuration of IDT hyperparameters found after 500 SMBO iterations

Hyperparameter	SMBO best value
SVR error penalty	2.51
Trajectory kernel width	0.151
HOG kernel width	0.223
HOF kernel width	0.010
MBH (x-direction) kernel width	0.657
MBH (y-direction) kernel width	12.55

Using this configuration an SVR model was trained using IDT descriptors extracted from the whole training set (1206 videos). After model fitting, test set exemplars were run through the model to produce perceptual load estimates, which are shown in comparison to the ground truth values in Figure 5-12 below.

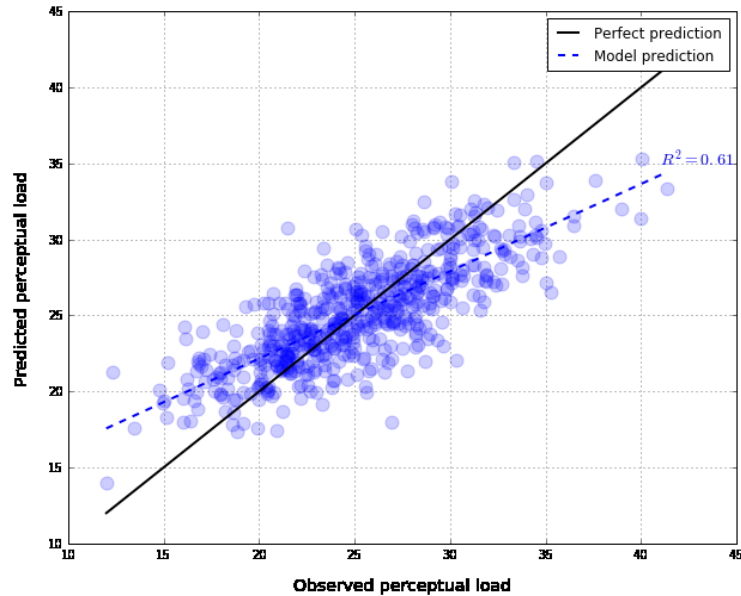


Figure 5-12. SVR performance of original IDT pipeline applied to the prediction of perceptual load. Each blue dot represents a test set exemplar

The original IDT descriptors are able to account for 60.7% of variation in held-out test set perceptual load values, a large increase over the performance of C3D, which obtained 56%. This may be explained by the fact that our IDT features are in fact partially learned from the data during codebook generation – as such they are more attuned to the distribution of motions and appearances seen in the data set – whereas we use a C3D model trained purely on a sports classification task. This suggests a future research direction concerned with retraining or fine tuning the C3D network specifically on a large corpus of driving videos to improve performance.

5.5.2 IDT+C3D with linear kernel

In the previous section, it was found that the IDT method outperformed C3D in representing spatio-temporal information relevant for perceptual load prediction; setting a benchmark on our dataset of 60.7% variance explained. In this and the following sections, it is therefore investigated whether the novel IDT+C3D combination can outperform these state-of-the-art methods in isolation. To set a baseline performance of the IDT+C3D descriptor, we first investigate regression performance when fusing the descriptors using a linear kernel.

Ridge regression. The ridge regularisation parameter was optimised using the TPE SMBO algorithm described above, which ran for 500 iterations. The progress of the SMBO algorithm in terms of best model performance over the 500 iterations is shown below.

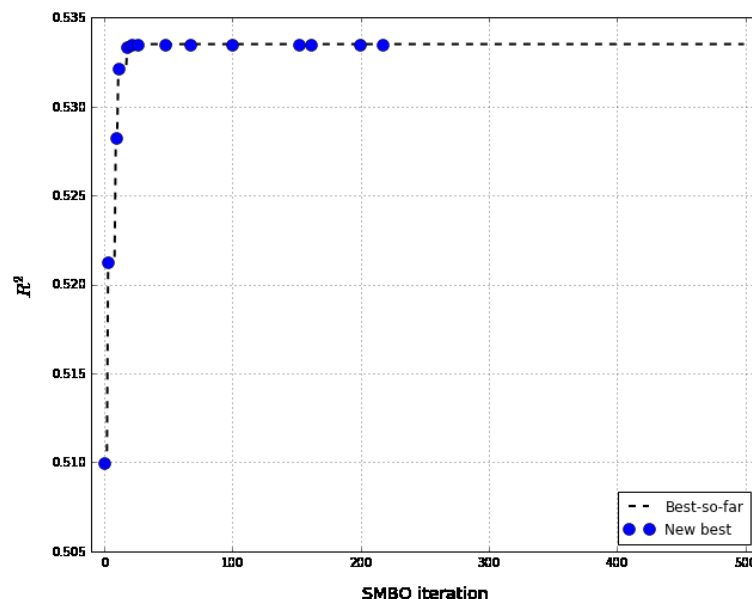


Figure 5-13. Progress of the SMBO algorithm in terms of model variance explained across the 500 algorithm iterations.

The best ridge regularisation parameter was discovered on the 218th SMBO iteration and corresponded to $\lambda = 8.97$, with an average cross-validation R^2 of .534 - rising from an initially sampled random regularisation model with an

average R^2 of .508. To estimate this configuration's generalisation performance, it was trained using the full (1206 example) dataset and perceptual load predictions were made on the held out validation set (603 examples); A model prediction vs. ground truth plot, displaying the true and predicted load value for each test set exemplar, is given in Figure 5-14.

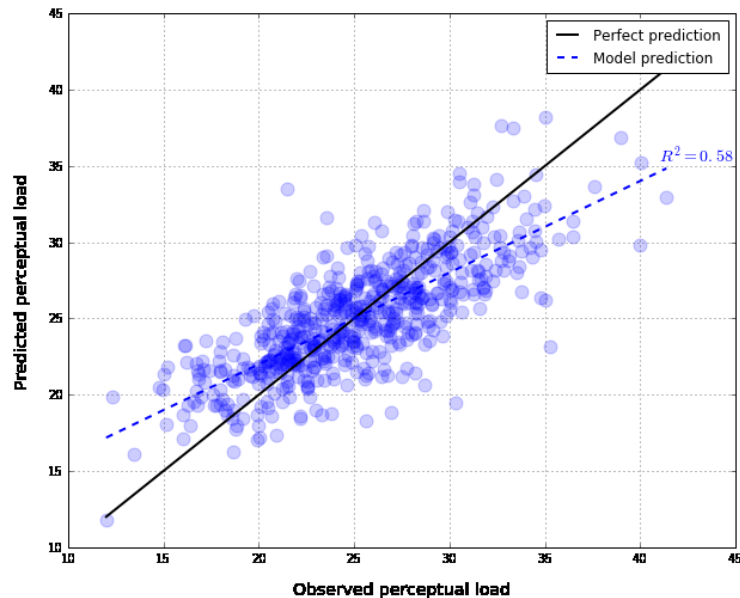


Figure 5-14. Ridge regression performance on held out test set for IDT+C3D descriptor combined using the linear kernel, where each blue marker represents a test set

The ridge regressor with regularisation parameter of 8.97 trained on the full training set produces a model which accounts for 58% of the variation in perceptual load values of the held out validation set ($R^2 = 0.58$).

SVR. The SVR C parameter was optimised using the SMBO algorithm, again with an initial sampling distribution defined as a lognormal distribution with mean 0 and standard deviation 1. After 500 iterations of SMBO the best performance was achieved with $C = 0.70$; the progress of SMBO is given in Figure 5-15.

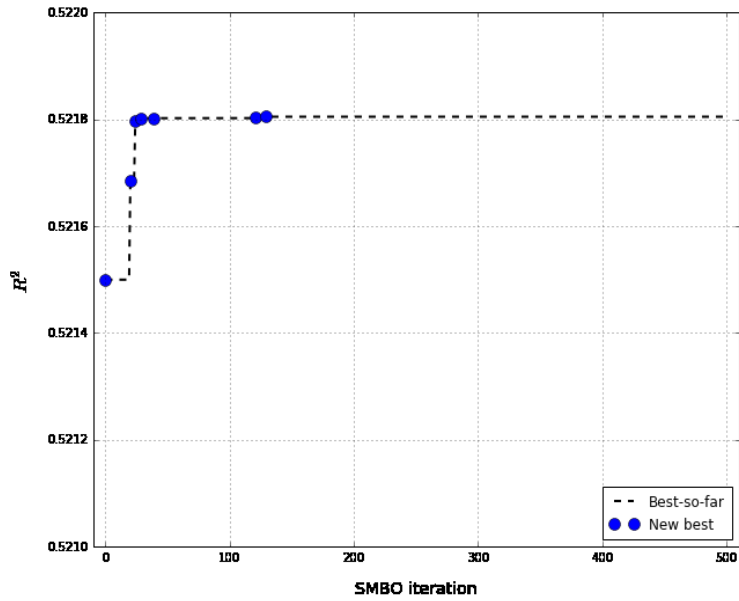


Figure 5-15. Progress of SVR with linear kernel channel fusion during 500 iterations of SMBO. Note the extremely small scale of the y-axis indicating the general robustness of SVR to the C parameter

Training this configuration on the full training set led to an R^2 value of 0.572 between the model's predictions and true load values of examples in the validation set. Figure 5-16 shows the predicted vs. true load values for each validation set exemplar.

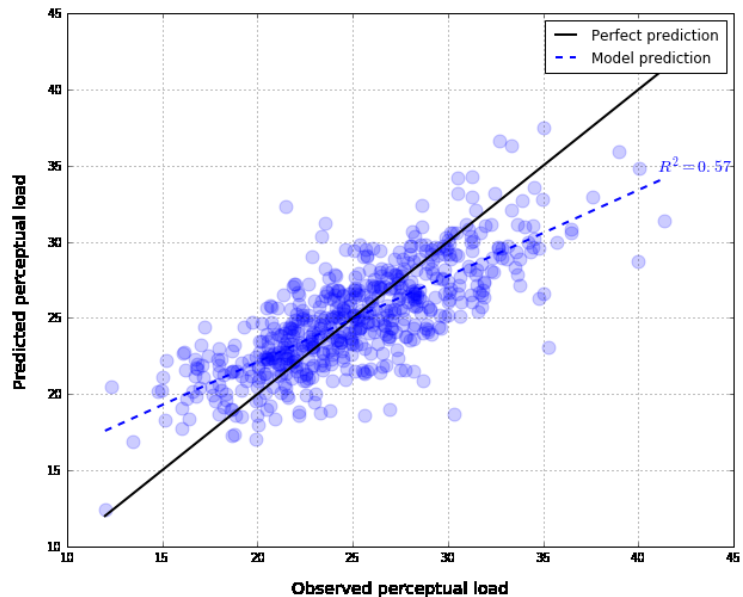


Figure 5-16. SVR with linear kernel, performance on the validation set after learning on full training set. Each blue marker represents a validation set exemplar

The linear SVR model's performance on the validation set is similar to that of the linear ridge regression, in that predictions are biased towards the mean, however overall performance is slightly reduced ($R^2 = 0.57$ vs. 0.58).

In summary then, it was found that a combination of IDT and C3D features does indeed encode a useful representation of perceptual load, even when a simple linear kernel channel fusion scheme is employed - more than 50% of variance was explained on a held out validation set by both SVR and ridge regression models trained on the 1206 example training set. However, performance was inferior to the original IDT technique in isolation, presumably due to IDT's use of nonlinearities in kernel mappings. Therefore, the next set of experiments investigates whether allowing nonlinear interdependencies between individual features improves the encoding of perceptual load by IDT+C3D to levels above the current state-of-the-art.

5.5.3 IDT+C3D with nonlinear kernels

In this section the performance of both SVR and ridge regression models are investigated while using a combination of χ^2 and RBF kernel functions to allow nonlinearity in the spatio-temporal features. For all configurations, each of the 5 IDT channels was kernelised using a χ^2 kernel and C3D with an RBF kernel. In an identical process to the last section, TPE-based SMBO routine was used to tune pipeline hyperparameters.

Ridge regression. Five χ^2 kernel width parameters, relating to the kernelisation of each IDT channel; a single RBF kernel width parameter, relating to the kernelisation of the C3D feature channel; and the ridge regularisation parameter were optimised using SMBO for 500 iterations. Once again, optimisation was

based on the mean R^2 across cross-validation folds on the restricted 1206 sample training set. The cross-validation R^2 progression through SMBO is shown in Figure 5-17.

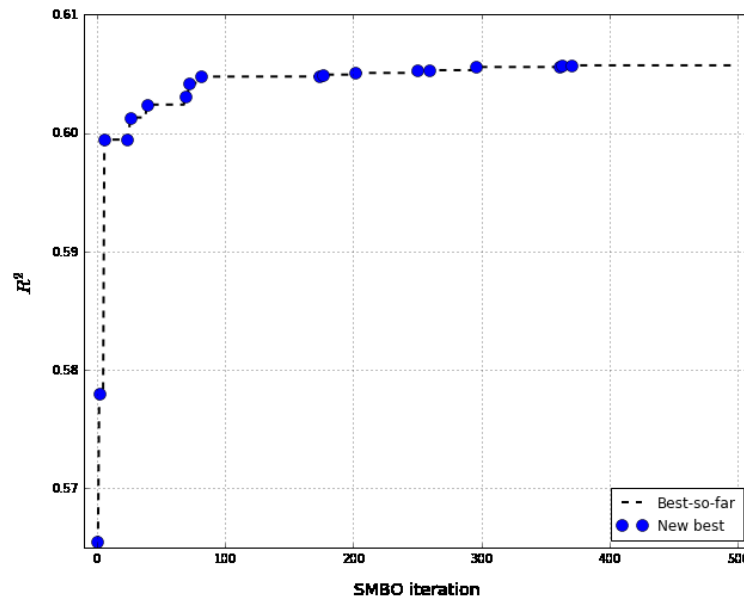


Figure 5-17. Progress of SMBO for ridge regression with nonlinear kernels on IDT and C3D features

Best performance was found on the 382nd iteration with $R^2 = 0.605$; this represents a large improvement over ridge regression utilising linear kernel representations ($R^2 = 0.534$) suggesting that nonlinearity in feature encoding uncovers information useful for predicting perceptual load from spatio-temporal features. The best hyperparameter values found by SMBO are shown in Table 5-4.

Table 5-4. Best found hyperparameters after 500 SMBO iterations for a ridge regression model with nonlinear multichannel kernel

Hyperparameter	SMBO best value
Ridge regularisation	0.121
Trajectory IDT kernel width	0.063
HOG IDT kernel width	0.117
HOF IDT kernel width	0.060
MBH (x-direction) IDT kernel width	0.180
MBH (y-direction) IDT kernel width	3405.482
C3D kernel width	0.793

Figure 5-18 displays a predicted vs. observed plot for a ridge regression model trained on the whole training set with the hyperparameters in Table 5-4.

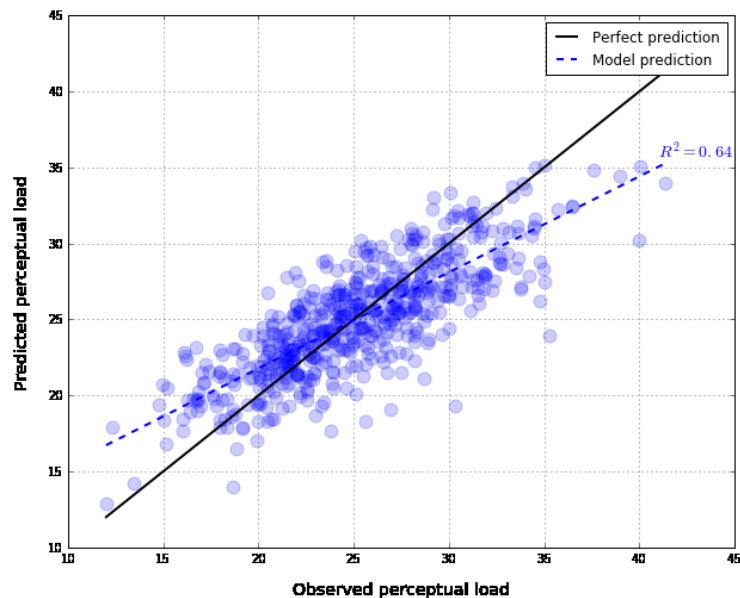


Figure 5-18. Predicted vs. actual load value plot for a ridge regression model using nonlinear multichannel kernel

Performance on the unseen validation set using the nonlinear kernels is again improved over the linear configuration ($R^2 = 0.637$ vs. $R^2 = 0.580$). Importantly, this also represents a substantial performance increase over IDT in isolation (R^2

= 0.637 vs. $R^2 = 0.607$), establishing a new overall best performance and indicating that the novel combination of IDT and C3D features combines information useful for perceptual load modelling. While the model is again biased towards mean load value, this effect is lessened in comparison to the linear kernel configuration; when nonlinearity in the mapping is made possible in the kernelisation of the features, a more accurate estimate of perceptual load is learned by the regression model. This confirms that there exist complex relationships between spatio-temporal features which carry information regarding the perceptual load of a driving scene; relationships which a simple linear mapping is unable to account for.

SVR. The SVR penalty parameter was optimised along with the 6 channel kernel width parameters using the same SMBO configuration as reported above for the ridge regression hyperparameter optimisation.

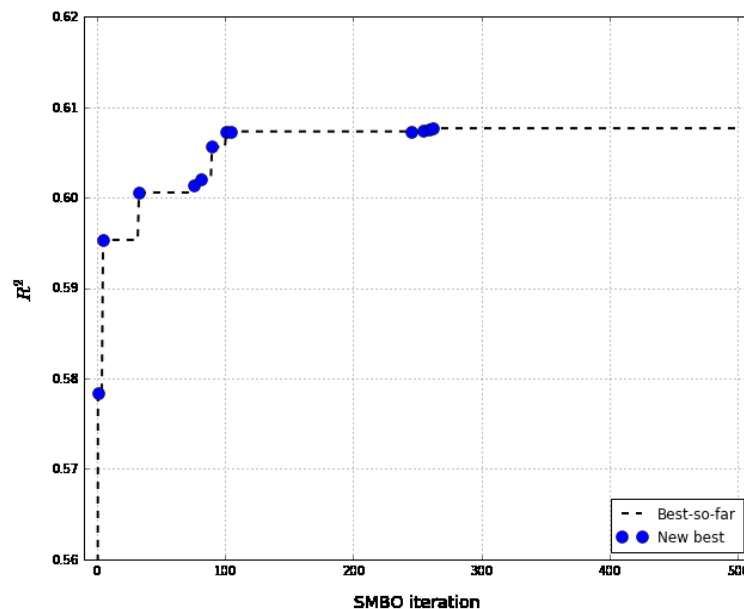


Figure 5-19. SMBO progression for SVR with nonlinear channel kernels

The best overall configuration was found on the 268th iteration, with a cross-validation R^2 of 0.607, again a higher performance than SVR using linear kernels ($R^2 = 0.522$). Table 5-5 displays the best found hyperparameters.

Table 5-5. Best hyperparameter set found with TPE-based SMBO for multichannel nonlinear kernel with SVR regression

Hyperparameter	SMBO best value
SVR penalty	5.612
Trajectory IDT kernel width	0.039
HOG IDT kernel width	0.105
HOF IDT kernel width	0.073
MBH (x-direction) IDT kernel width	0.141
MBH (y-direction) IDT kernel width	18.297
C3D kernel width	0.744

The best found kernel hyperparameters for the SVR model are closely aligned to those found for the ridge regression previously, suggesting a robust local minimum around these kernel widths. After training this model configuration on the whole training set a validation set R^2 of 0.626 was achieved, as seen in Figure 5-20, a slight decrease in performance in relation to the previous variant (using ridge regression).

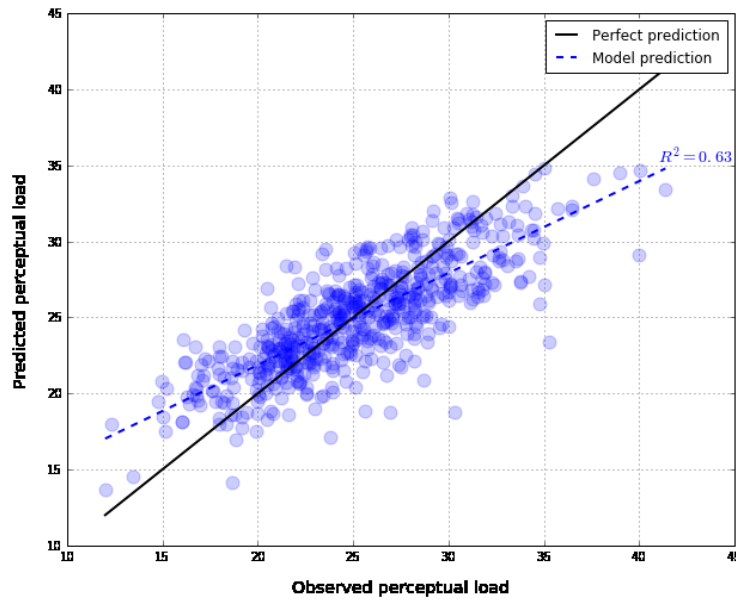


Figure 5-20. Predicted perceptual load values of validation set examples plotted against actual values for the nonlinear multichannel kernel SVR model

Ridge regression with channel weights. The model is now expanded such that individual channel weights of the multichannel kernel become a free parameter. A channel weight, W^c , is now introduced into the kernel function, such that

$$K(x_i, x_j) = \sum_c W^c k^c(x_i^c, x_j^c) / A^c$$

For the investigation into channel weights, the associated kernel widths were fixed at the best configuration found for nonlinear ridge regression without channel weights. An SMBO routine was therefore employed to optimise the 7 hyperparameters: the ridge regularisation parameter, again with an initial lognormal sampling distribution with mean 0 and standard deviation 1, and each channel's weight in the multichannel kernel computation, each of which was assigned a uniform sampling distribution with a minimum of 0 and maximum of 1. The TPE-based SMBO was again run for 500 iterations, the progress of which can be seen in Figure 5-21.

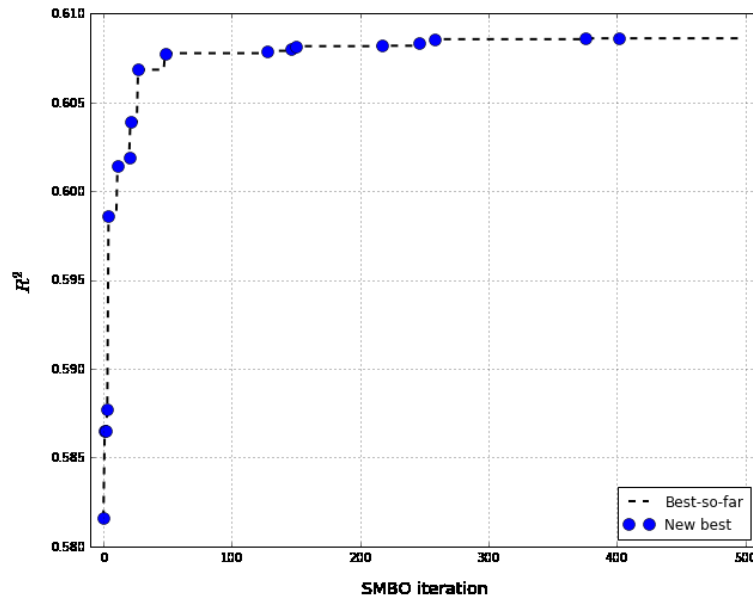


Figure 5-21. SMBO progression over 500 iterations for optimising regularisation and channel weights for multichannel kernel computation and ridge regression

The best performing configuration of channel weights and regularisation parameter was found on iteration 401 of the SMBO routing, resulting in a cross-validation R^2 of 0.608, slightly outperforming the constant weight ridge regression variant investigated earlier. The best hyperparameters are reported below.

Table 5-6. Hyperparameters optimised after 500 iterations of SMBO for nonlinear kernel ridge regression with variable channel weights.

Hyperparameter	SMBO best value
Ridge regularisation	0.141
Trajectory channel weight	0.543
HOG channel weight	0.530
HOF channel weight	0.261
MBH (x-axis) channel weight	0.998
MBH (y-axis) channel weight	0.109
C3D channel weight	0.554

The channel weights can be interpreted as a form of feature importance. It is clear that the motion boundary histogram (MBH) channel in the x-axis is weighted highly by the model: this makes intuitive sense as motions of objects moving across the visual field of the driver in the horizontal directions represent occurrences that will often require attending to another location, for example a pedestrian approaching the road from pavement, or a car approaching an intersection from a perpendicular road. This is most starkly seen in the comparison with MBH channel in the y-axis, which has a relatively low channel weight, which rarely occurs during driving and would not likely indicate a trajectory of an object intersects the ego-car's motion. A likely reason for the histogram of optical flow (HOF) channel being suppressed is that much of the relevant information contained within is represented more clearly in the MBH channels, specifically motion, as the HOF feature does not take into account the motion of the ego-car as the MBH feature does.

Training a ridge model with the above hyperparameters resulted in an R^2 score of 0.629 on the unseen validation set, slightly lower than the uniformly weighted ridge regression variant reported earlier.

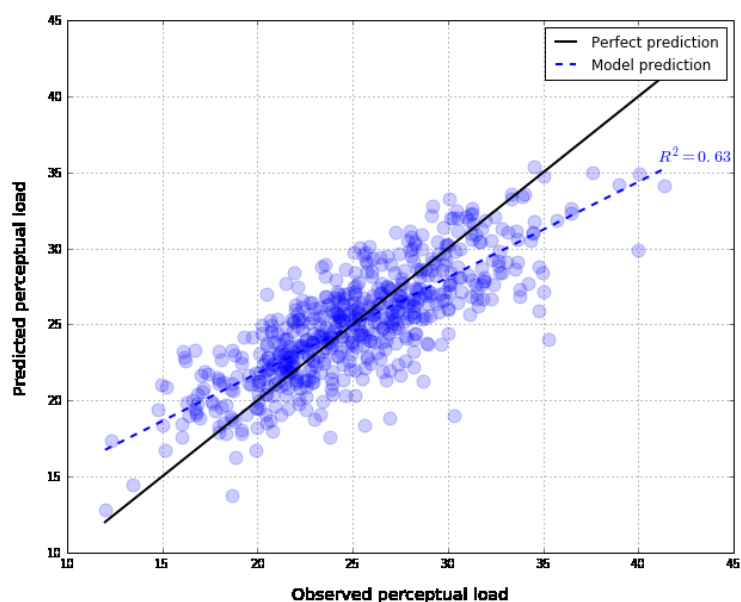


Figure 5-22. Best performing configuration of ridge regression with individual channel weights on the held-out validation set

5.5.4 IDT+C3D results summary

Table 5-7 contains a summary of the best performing configurations found using the TPE-based SMBO algorithm for each of the model and kernel configurations explored for the IDT+C3D descriptor.

Table 5-7. R2 scores of the best performing models in each kernel-regression configuration during: SMBO cross-validation (in columns labelled CV) and performance on the unseen validation set (in columns labelled Val). Best cross-validation and validation scores across all models are in bold

Kernel →	Linear		Nonlinear		Nonlinear (with channel weights)	
	CV	Val	CV	Val	CV	Val
SVR	0.522	0.572	0.607	0.626	N/A	N/A
Ridge	0.534	0.580	0.605	0.637	0.608	0.629

Importantly, for all nonlinear fusion schemes it was shown that the IDT+C3D descriptor outperformed both IDT ($R^2 = 0.607$) and C3D ($R^2 = 0.558$) in isolation on the held-out test data. Another clear finding is the superiority of nonlinear kernels in capturing information relating to perceptual load in comparison to linear kernels, for both SVR and ridge regression models, implying a complex mapping between features is necessary for characterising perceptual load in driving scenes. Furthermore, ridge regression was shown to better capture the variance in perceptual load for both linear and nonlinear multichannel kernelisation schemes. Introducing individual channel weights as free parameters of the model allowed the investigation of feature relevance, confirming intuitions regarding the relative importance of movement directions in the scene. Performance on the held-out validation set was slightly reduced in comparison to the equal channel weight variant, perhaps due to an interaction

between channel weights and kernel widths, although cross-validation performance during SMBO was improved.

5.6 Chapter Discussion

The principal finding of the chapter establishes that it is possible to robustly predict the perceptual load of a driving situation using only the visual information contained in a video depiction: after an exhaustive survey of system configurations, the best performing variant was able to account for 63.7% of the variance in perceptual load across an unseen set of videos. In contrast to previous attempts at estimating perceptual load *a priori* from the visual attributes of a task (e.g. Roper et al., 2013; Dayan & Solomon, 2010), our method predicts load in a dynamic real-world task defined in terms of complex natural imagery. Furthermore, the system here produces estimates of perceptual load from raw pixel data through automated descriptor extraction and regression processes rather than through recourse to human-defined intermediate categorisation of the visual stimuli (e.g. the imprecise ‘similarity’ concept used by Roper et al., 2013). As such, the work here represents the first objective, reproducible, method to estimate load in a real task using only the visual information of the task, without any human intervention in the process.

This was made possible by framing perceptual load as a subjective attribute of the driving scene. Through collecting more than 50,000 pairwise judgements between video clips depicting different driving situations, ground-truth labels of perceptual load were assigned to each of 1809 collected video clips by the application of the TrueSkill rating algorithm (Herbrich et al., 2006). Analysis showed that these labels were stable before all comparisons were counted, indicating robust and consistent labelling across comparisons. The construction

of this large video corpus and associated labels constitutes a novel application of subjective attribute measurement methods, which have so far been predominantly applied to measuring attributes within images rather than video (e.g. Donahue & Grauman, 2011; Kiapour et al., 2014) and with more directly ‘visual’ attributes, such as the shininess of shoes (Kovashka et al., 2012) or clothing style (Yamaguchi et al., 2015); although recent work has also investigated image ‘interestingness’, which may be more conceptually similar to perceptual load. One criticism which could be levelled against the work however, is that what we have measured and subsequently modelled is not perceptual load *per se* but in fact the *perception* of perceptual load, were one to be in that driving scenario. This is a valid concern, and one which naturally suggests further validation experiments to confirm that, for example, detections of critical stimuli are reduced when the model’s estimate of perceptual load is high. However, given previous work in related domains showing that similar subjective assessments indeed correlate with objective measures (e.g. task-load; Mazur et al., 2013) it is reasonable to assume that our TrueSkill estimates do indeed correlate with actual perceptual demand to a real degree.

The work also introduced a novel hybrid video descriptor, IDT+C3D, a fusion of existing improved dense trajectories (IDT; Wang et al., 2013; 2015) and 3D convolutional neural network (C3D; Du Tran et al., 2014) approaches to video description. Both methods in isolation have previously reported state-of-the-art performance on several human action recognition challenges, and so were chosen here to characterise the rich motion and appearance patterns present in urban driving. Both were benchmarked on our perceptual load in driving dataset, where IDT outperformed C3D in predicting perceptual load for unseen test set videos. However, through casting C3D as a separate channel of information, and combining with IDT features in a multichannel kernel, much improved perceptual load estimates were observed. The power of our novel

IDT+C3D descriptor may lie in the different biases in representational content in IDT and C3D individually: while IDT captures long and complex motion patterns and associates them with relatively basic appearance descriptors (e.g. HOG; Dalal & Triggs, 2004), C3D learns especially rich appearance representations, due to its deep feature learning convolutional architecture, which are shallow in time. The C3D channel therefore contributes complementary spatial appearance information to the IDT channels, resulting in an improved dynamic scene representation more suitable to capturing the visual variations relevant for perceptual load.

6 General discussion

6.1 Summary of findings

The aim of this thesis was twofold: 1) to establish neural mechanisms for widely observed perceptual deficits induced by high perceptual load tasks on perceptual systems, and 2) to create a predictive model of the perceptual demands of a visual task using the visual information available in the task itself. In Chapters 3 and 4, recent neuroimaging methods were employed to measure the neural excitation associated with oriented visual stimuli while participants completed tasks of both low and high perceptual load. Chapter 5 expanded and developed modelling approaches rooted in computer vision and machine learning to estimate perceptual load in the dynamic real-world task of driving.

In Experiment 1 of Chapter 3, deleterious effects of high perceptual load were observed on orientation change detection using novel experimental stimuli and psychophysical design, consistent with previous findings of reduced stimulus detection under high load (e.g. Macdonald & Lavie, 2008; Carmel et al, 2007). While perceptual load was modulated with a primary RSVP task at fixation, orientation change detection accuracies were measured across ranges of orientation change offsets in a secondary peripheral task. It was found that high load led to an overall suppression of change detection accuracy across the offset range and an increased orientation offset threshold for detection, thus confirming perceptual load effects on the perception of oriented grating stimuli. In Experiment 2, multivariate pattern analysis (MVPA) methods were used to quantify the representational content of visual areas V1, V2, and V3, in response to large oriented gratings under low and high perceptual load. While univariate BOLD signal analysis confirmed a general suppression of activity in

cortical areas, no statistical difference was found by MVPA methods between pattern classification accuracy for orientation-specific activity patterns elicited under low and high load conditions. This unexpected result is perhaps best explained by properties of the experimental design: specifically, participants' spare attentional capacity under low load may not have been allocated to perceiving the orientation of the peripheral grating due to attentional capture by other properties of the stimulus, and furthermore altering the range of oriented gratings to enable the construction of voxel tuning function (VTFs), which characterise the tuning properties of neural populations explicitly.

The design of the experiment in Chapter 4 was therefore modified to address these possibilities. A secondary delayed orientation discrimination task was included to ensure spill-over of resources to the perception of orientation, while the orientation range of the grating stimuli was extended across the full range to allow the construction of voxel tuning functions (VTFs). Perceptual load was again manipulated with a central primary task, performance on which was reduced in the high load condition indicating the effectiveness of the load manipulation. Performance on the secondary orientation discrimination task was high under both load conditions (more than 80% detection accuracy for 20° offsets), suggesting that perceptual resources were indeed directed to the perception of the gratings, and furthermore performance was decreased when the primary task load was high. Analysis of BOLD signal again found a general suppression of visuo-cortical activity across early visual cortex in the high load condition, replicating previous findings of reduced visual processing under load (Rees et al., 1997; Schwartz et al., 2005).

To investigate orientation-specific modulations, VTFs were then constructed using responses elicited by the gratings across V1, V2, and V3: while no statistical difference was found between VTFs extracted for V2 and V3 voxels,

tuning functions constructed from V1 activity were found to have *both* reduced amplitude and increased bandwidth under high perceptual load relative to low. This finding suggests a novel mechanism of action for load-induced perceptual deficits which originates at the very earliest stages of cortical processing.

Interestingly, SVM-based MVPA using distributed patterns of activity in Chapter 4 also showed no significant difference for orientation classification accuracy between activity elicited under high and low load in any of the investigated visual areas (V1, V2, and V3). This apparent discrepancy between VTF and MVPA findings may be resolved by the fact that MVPA is a representational measure across sets of voxels whereas VTF analysis characterises the tuning properties of neural populations within individual voxels. It does not follow then that MVPA accuracy is always positively correlated with VTF multiplicative or bandwidth scaling, for example in the case where the distribution of voxel orientation preferences is non-uniform. This possibility was addressed, where it was found that orientation preferences across voxels were not uniform across the probed orientation range, voxel preferences being biased towards the horizontal meridian. The finding suggests that perceptual load acts to degrade orientation encoding at the level of local neural populations in cortex rather than across whole retinotopic areas.

In Chapter 5 a predictive model of perceptual load was produced using computer vision and machine learning techniques. A dataset of 1809 short video clips was collected depicting real-world driving and a value of perceptual load assigned to each clip through aggregating more than 60,000 pairwise comparisons between clips, collected in a large-scale experiment. Spatio-temporal features, used successfully in previous work to classify motion-based actions from video (i.e. improved dense trajectories; Wang et al., 2013; 2015; and convolutional 3D features; Du Tran et al, 2015), were extracted from the

video clips to produce parsimonious clip representations. IDT and C3D representations were then fused using a multichannel kernel and mapped to the perceptual load values derived from pairwise comparisons using regression analyses. A variety of methods for feature fusion and regression were compared: it was found that non-linear feature fusion, utilising a combination of all IDT and C3D features, produced improved regression performance over linear methods, indicating that complex interactions between motion features across the visual field are informative for the prediction of perceptual load. The best performing model configuration resulted in 63.7% of variance in perceptual load values being explained by the model. Furthermore, an investigation into the relative importance of certain features in the model's accuracy found that trajectory motions in the x-axis (i.e. horizontal across the visual field) were relatively more useful for the model, confirming intuitions regarding such motions during driving (e.g. pedestrians crossing the road). The system therefore constitutes the first model to predict, *a priori* from visual information, the perceptual load induced by a complex, dynamic, real-world task.

6.2 Perceptual load and orientation encoding

6.2.1 The unique effect of perceptual load on orientation tuning

The findings of Chapters 3 and 4 suggest that perceptual load degrades the perception of orientation, through a novel mechanism which can be traced to a change in processing at the earliest stages of orientation representation. While previous studies (e.g. Rees et al., 1997; Schwartz et al., 2005) have found an overall, feature-independent, reduction in visuo-cortical activity to visual stimulation under high perceptual load, the findings presented here show that load also modulates *feature-dependent* tuning of neural populations responding

to orientation through a novel combination of mechanisms. Low-level orientation tuning curves can be altered in a feature-dependent fashion in two ways: 1) multiplicative scaling, where activity of the neural population is scaled linearly with respect to the proximity of the viewed orientation to the population's preferred orientation (i.e. a change in response profile amplitude); and 2) bandwidth scaling, where the population's response becomes more selective to the preferred orientation (i.e. a change in response profile width). Through VTF analysis in Chapter 4, we show that perceptual load induces both multiplicative and bandwidth scaling, along with feature-independent additive activity shift, in populations of V1 neurons. Under load orientation response profile amplitude is reduced, width is increased, and overall activity is suppressed.

This combination of effects is novel in the literature, where previous research has investigated the modulations of feature-specific tuning physiologically and using fMRI methods for single cells and neural populations in the context of several attentional manipulations; no previous manipulation has reported finding these three effects in parallel. At the single cell level, it has been reported that modulations of tuning curves are restricted to multiplicative scaling of responses across the curve, a result which has been shown for orientation tuned cells in V4 (McAdams & Maunsell, 1999), direction tuning in MT (Treue & Martinez-Trujillo, 1999), and contrast response functions in MT (Lee & Maunsell, 2010). Cueing of attention to a specific location in the visual field is known to additionally induce an additive shift in activity for single cells and neural populations with receptive fields overlapping the cued location. This effect is present for neurons in the early visual cortex of macaques (Moran & Desimone, 1985; Luck et al., 1997) and in neural populations of human visual cortex using fMRI (Kastner et al., 1998; Saproo & Serences, 2010).

These results suggest that *bandwidth* scaling does not occur during the allocation of spatial attention, however Martinez-Trujillo and Treue (2004) subsequently investigated the responses of single cells and neural populations in macaque V4 while manipulating the allocation of feature-based attention. They found that orientation tuning curves extracted for attended stimuli were more selective (i.e. showed reduced bandwidth) in comparison to unattended stimuli, while no multiplicative or additive scaling was observed. This result was specific to neural populations, however: tuning curves constructed using single cell activity showed only a multiplicative scaling for attended stimuli, with bandwidth scaling on the population level being an emergent property across neurons. Therefore, previous attentional manipulations have induced all three types of response scaling, however never in unison by a single manipulation. The present work therefore suggests a unique role for perceptual load in shaping low-level feature representations in primary visual cortex in the case of orientation.

6.2.2 Perceptual consequences of load-induced modulations

While these tuning effects are indeed novel, it is useful to expound upon the consequences of such changes to perception; can they account for the perceptual deficits associated with high load? With regard to feature-independent activity shift, the effect of a global suppression of visuo-cortical activity (as reported in Chapter 4 and Experiment 2 of Chapter 3) is to reduce the signal-to-noise ratio for stimuli presented outside the central focus of attention (Mangun, 1995). In the context of a biased-competition account of perceptual processing (Desimone & Duncan, 1995), this reduction leads to a relative attenuation in the transmission of low-level stimulus characteristics to higher levels of perceptual analysis. The observed reduction of signal in visual

cortex under load therefore, in and of itself, contributes to deficits in higher-level perception such as object recognition or stimulus detection for task-irrelevant stimuli.

To understand the perceptual effects of feature-dependent activity modulations, such as multiplicative scaling (i.e. tuning curve amplitude change), and bandwidth scaling (i.e. tuning curve width change), it is useful to refer to the amount of information a population of neurons tuned to a given feature is able to convey regarding the state of that feature (Panzeri et al., 2008). Mutual information (MI) is an information-theoretic measure, which conveys the reduction in uncertainty of one variable conditional on knowing the state of some other variable. In the current context, MI would relate the BOLD activity of a voxel to the reduction in uncertainty regarding the orientation of the presented stimulus; for example, a population of neurons with high MI would encode stimulus orientation with relative certainty. Sprague et al. (2015) showed mathematically that positive multiplicative scaling always increases the MI between neural activity and stimulus state. Therefore, our finding of reduced amplitude of V1 population tuning curves under conditions of high load implies that V1 orientation response profiles contain less information regarding the actual stimulus orientation. This in turn indicates that visual areas more advanced in the visual hierarchy, which take inputs from such V1 feature detectors, receive less information regarding the state of oriented edges in the visual scene, thus hindering the formation of percepts useful for higher-order operations such as object detection and recognition (Ditterich et al., 2003; Shadlen & Newsome, 2001).

The relationship between tuning curve bandwidth and mutual information is non-monotonic, however (Series et al., 2004). A decrease in bandwidth would increase MI if the bandwidth was originally suboptimal (e.g. in the case of a flat

tuning function). However, further reductions in bandwidth beyond the optimal level would lower MI, as the tuning function would convey information about a very restricted range of stimulus values. Thus an increase in bandwidth - to a point - would help if the original bandwidth was overly narrow. The optimal population tuning width is not solely dependent on the overall shape of the tuning curve however, but also on the type of neural response noise, and the prior distribution of orientations in the domain (Sprague et al., 2015). Therefore, in terms of direct estimation of orientation from a population response, the effect of V1 bandwidth scaling under load is difficult to estimate exactly, although given the wide bandwidths observed for V1 populations in Chapter 4 it is likely to be deleterious.

The utility of low-level orientation tuning may not only be in terms of direct estimation of the stimulus orientation, however. For example, performance in the task of discriminating two oriented stimuli does not follow the same dependencies on tuning curve shape as the estimation of absolute orientation. Kang et al., (2004) have shown that the optimal population orientation tuning width in a discrimination task is linearly dependent on the angular offset between the stimuli to be discriminated, such that the optimal width of a tuned population for a given orientation discrimination is approximately 0.3 times the offset magnitude. The optimal tuning width, while varying with orientation discrimination offset, was theoretically shown to be below 20° (measured as tuning curve full-width half maximum) for all possible offset magnitudes: given the widths of tuning curves extracted under low and high perceptual load conditions in Chapter 4, in both cases being larger than 20° , the observed bandwidth scaling indicates that the ability of those populations to discriminate between any pair of oriented gratings is reduced under load. This result was indeed observed behaviourally in Experiment 1 of Chapter 3 and the secondary discrimination task of Chapter 4.

Given that extraction of orientation information is a fundamental step in the hierarchy of visual processing, the novel scaling effects therefore suggest that orientation-specific information transmitted from V1 to higher areas in the visual hierarchy is less precise under load. As this information forms the basis for the representation of higher-level visual concepts, such as objects, the observed response profile scaling provide a new mechanism to explain previously reported high-level effects of perceptual load, such as load-induced inattentive blindness and change blindness (e.g. Macdonald & Lavie, 2008; Carmel et al., 2007). The scaling effects are also consistent with more recent findings in the psychophysics and fMRI domains. For example, Stolte et al (2014) psychophysically obtained orientation response curves using an orientation masking paradigm, finding that tuning curves of discrimination accuracy constructed across contrast levels were significantly widened under high perceptual load. Their result can now be explained in terms of the scaling effects of population tuning in primary visual cortex, rather than through recourse to higher order processes which occur between early visual processing and behavioural response during a psychophysical experiment. Our result is also consistent with work by de Haas et al. (2015) who found modulations of spatial tuning in early visual cortex due to perceptual load. They reported an increase in width for population receptive fields and a shift in field center under load, which together with the current findings indicate an important role for task-related perceptual load in shaping visuo-cortical response to fundamental features of the visual field.

6.3 Modelling load

6.3.1 Relation to other models of perceptual load

The model of perceptual load constructed in Chapter 5 represents two major steps forward for the theory of perceptual load. It constitutes the first application of perceptual load theory to the rich content of natural scenes as well as being the first attempt to model perceptual load in a non-laboratory task. In this section I discuss how the model relates to previous efforts to model perceptual load. The most directly relevant previous attempts at creating predictive models of perceptual load level were those of Roper et al. (2013) who employed a data-driven regression framework, and Dayan and Solomon (2010) who used a Bayesian approach. Both models were defined only for the laboratory-based response competition paradigm however (Eriksen & Eriksen, 1974; Lavie, 1995), and attempt to predict flanker interference effects under load; here I describe each in turn.

Although the work reported in Chapter 5 is a novel approach to the study of perceptual load, there exist parallels to the work of Roper et al. (2013), who attempted to estimate flanker effects due to load in response competition using visual information and performance in an independent task. With a linear regression analysis they produced a model including regressors for target and distractor similarity combinations as well as search slope and intercept for a visual search task using the same display. While their model produced good predictions of flanker effects, they somewhat surprisingly found that high visual complexity of the display in and of itself, as defined by low target-distractor similarity in combination with low distractor-distractor similarity, did not lead to a good level of prediction of exhausted attentional capacity and the associated inhibition of flanker interference. Rather, performance on the independent visual search task was found to be the main predictor of perceptual load effects. For this reason, they cautioned that it may not be reasonable to classify task displays based on phenomenology alone, suggesting that there is a need for an

independent measure, such as visual search performance, to estimate the perceptual load induced by a visual task. However, the model described in this thesis is successful at estimating perceptual load values using only visual information, without recourse to additional independent measures. Their suggestion does however present a promising line of future work in incorporating other, perhaps either behavioural or physiological, markers of load into the model as regressors.

This difference in conclusion and attitude towards using solely visual characteristics of the task to estimate load may be attributable to the level of information extracted from the displays in Roper et al.'s (2013) model versus ours. Roper and colleagues hand design and categorise visual displays by intuitive label combinations, such as “low target-distractor similarity”, and build the model using these labels as the representation of visual information in the field. In our case, primitive visual features of the scenes and their transformation through time (IDT; Wang et al., 2013, 2105; C3D; Du Tran et al., 2014) were extracted from the 'displays' (i.e. video clips). While these descriptors were chosen as they have been shown to capture useful information regarding moving objects for classification, they were not based on human judgements on higher-level constructs such as object similarity. As such, the model presented here has more freedom to construct abstract informative features from the initially extracted primitives.

Employing a Bayesian approach to modelling perceptual load, Dayan and Solomon (2010) also produced a mathematical explanation of load-induced flanker effects in the response competition paradigm. In their formulation, the visual elements of the task were represented with a primitive binary coding scheme, where display elements (i.e. targets, flankers, and distractors) were represented as being present or absent at specific locations, while simple model

neurons were configured to respond to the presence of display elements within their receptive field. The model was able to account for flanker effects due to the competition amongst receptive fields for representation (a possibility also explored by Scalf et al., 2013), however their approach, as is, could not account for a more realistic version of the task, given the simplistic nature of visual information used as input for the model. This criticism is especially so for tasks involving rich real-world imagery where even the notion of such a representation is not readily available; in that defining relevant object boundaries and relevant task units *a priori* for a given task is itself not well understood.

6.3.2 Contributions to computer vision

Many applications of computer vision involve the prediction or estimation of classes or values where the visual properties are objective – that is, there is little ambiguity in the true value of the property for a human labeller. For example, much research has concerned predictive models of object category. For this purpose, large ground-truth datasets are produced by humans annotating images, indicating whether they contain a certain object class, such as 'dog' or 'car'. Typically human annotators display high levels of inter-annotator reliability for objective classifications such as these (Nowak & Ruger, 2010). The related but distinct problem of estimating subjective attributes (e.g. how 'gothic' a style of clothing is; Kiapour et al., 2014) from visual information is a more recent development, however (e.g. Farhadi et al., 2009; Wang & Mori, 2009; Berg et al., 2010). The difficulty inherent in modelling such attributes lies in their subjective nature: what constitutes gothic clothing to one annotator may not be to another. This difficulty is exacerbated when annotators are asked to produce absolute values for attributes on a scale, for example, what does '5 out of 10' mean in terms of gothicness?

It has been shown that people produce more reliable estimates for such attributes when asked to compare images and rank them with respect to the attribute of interest (Ma et al., 2012). This approach has therefore been adopted widely in the recent attribute estimation literature, for: estimating facial attractiveness (Donahue & Grauman, 2011), estimating properties of consumer goods such as shoe shininess (Kovashka et al., 2012), and estimating image interestingness, which has seen much recent research activity (e.g. Dhar et al., 2011; Gygli et al., 2013). The work presented in Chapter 5 therefore constitutes a novel application of these methods, in estimating the psychological phenomenon of perceptual load. Through casting perceptual load as a subjective attribute of a scene, and estimating ground-truth load values from many relative pairwise rankings, the prediction of perceptual load could be represented by a regression problem from video features to perceptual load. The work also represents the first principled attempt at subjective attribute estimation in the spatio-temporal domain of video. While Jiang et al. (2013) attempted to estimate the interestingness of videos, their method of obtaining ground-truth interestingness relied on the in-built search functionality of Flickr (which has a 'sort by interesting' feature); the mechanism of this algorithm is unknown and therefore the values obtained should be treated with skepticism. Our bottom-up approach to obtain attribute estimates using a pairwise comparison methodology is therefore unique in the video analysis literature.

The model presented in Chapter 5 also constitutes the first application of *spatio-temporal* representations to estimate attributes in video. Previous approaches (e.g. Jiang et al., 2013; 2014) have relied on frame-level features for video representation, such as HOG (Dala & Triggs, 2004), SIFT (Lowe, 2000), and GIST (Torralba et al., 2006). We utilised improved dense trajectories (IDT; Wang et al; 2013; 2015) and C3D (Du Tran et al., 2014) which have shown

state-of-the-art performance in generating semantically meaningful video representations in related domains. Furthermore, it was shown that combining both descriptors, to produce a novel IDT+C3D representation which combines the motion-oriented representation of IDT with the excellent appearance representation of C3D, showed improved perceptual load prediction performance in comparison to each approach individually. The method developed for fusing IDT and C3D representations also represents a novelty, where deep convolutional neural network (C3D) and handcrafted (IDT) features were aggregated through the use of a multichannel kernel. Although deep convolutional neural networks have been shown improved performance through the incorporation of additional handcrafted features, or by fusing predictions from multiple CNNs, these combinations are usually implemented via feature concatenation (e.g. Laptev et al., 2008; Wang et al., 2015), or by averaging output predictions from several models or network streams (e.g. Tang et al., 2013; Acar et al., 2015). In Chapter 5, extending the multichannel approach employed originally by Wang et al. (2013), through incorporating deep convolutional features containing rich appearance and short-term temporal information, enabled the construction of a compact, nonlinear, kernel-based representation of perceptually relevant features which significantly outperformed both IDT and C3D methods in isolation.

6.3.3 Applied implications

The construction of a robust model of perceptual load in Chapter 5 for the complex real-world task of driving has clear practical applications in automotive technologies, as well as for safety-critical environments in general. Previous applied work perceptual load in driving has until now focused on the *effects* of increased perceptual load, whether originating from inside (e.g. a secondary

task to complete whilst driving) or outside (e.g. the arrangement of vehicles) the vehicle, on driver performance. For example Yeshurun and Marciano (2015) varied the number of pedestrians or vehicles in a driving simulator, as proxies for perceptual load, and recorded driving performance measures such as speed and number of crashes. Similarly, Murphy and Greene's (2016) simulator study varied perceptual load in two levels – by making a gap for the subject's car to fit through either tight (high load) or with plenty of room (low load) – and recorded whether the subject detected a critical stimulus, such as a pedestrian standing to the side of the road. While such approaches are important for establishing the relevance of perceptual load in complex tasks such as driving, given that the manipulations of load are pre-determined and rendered in driving simulators, there does not seem a natural way to take the findings from the laboratory to the real-world task of driving, where the visual scene is not under experimental control.

The approach developed in Chapter 5, which maps directly from the *actual* visual state of a *natural* task (rather than using predetermined task arrays or conditions) to an estimate of induced perceptual load, therefore constitutes a useful and implementable system for real-world driving today. Given that visual information present during a visual task is easily captured with affordable, high-definition, small-size cameras, and given a pretrained model of perceptual load in terms of properties of the scene, the visual scene itself could be used to detect high levels of perceptual load and therefore inform warning systems of safety or time-critical operator's (e.g. drivers, pilots, astronauts) reduced ability to detect critical stimuli. Unlike relying on secondary-task performance, this measure can index the level of ongoing perceptual load without imposing any additional demands and without interfering with the operators' work. Therefore, estimating perceptual load directly from live dynamic imagery has the potential to become instrumental in managing perceptual load in daily-life safety-critical

activities such as driving, as well as in industries ranging from aviation to healthcare.

6.4 Concluding remarks

Failures of vision such as inattention can have profound consequences for our ability to act in our environment. The work presented in this thesis aimed to uncover the neural mechanisms by which increased perceptual load contributes to such failures, whilst also providing a means to estimate the magnitude of this critical characteristic in complex real-world tasks. The findings extend the scope of perceptual load theory, indicating that perceptual load degrades the encoding of the visual feature of orientation, a fundamental building-block of visual perception, in unique ways at the very earliest stages of cortical processing. Through reduced orientation selectivity in early visuo-cortical neural populations, as well as the overall magnitude of neural response, perceptual load diminishes the fidelity of low-level representations of orientation which are crucial in forming coherent higher-level percepts; the work thus reports a novel mechanism for the characteristic failures of vision seen under load. While load was manipulated experimentally to uncover this novel mechanism, the thesis also aimed to develop a method for predicting the level of perceptual load in complex natural visual tasks. Through a combination of computer vision and machine learning techniques, a model was developed for the dynamic task of driving which robustly agrees with human judgements of attentional demand during urban driving. The model, and indeed the adaptable modelling method itself, represent major steps towards applying the science of perceptual load theory to critical, real-world, situations.

References

Acar, E., Hopfgartner, F., & Albayrak, S. (2015, June). Fusion of learned multi-modal representations and dense trajectories for emotional analysis in videos. In *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)* (pp. 1-6). IEEE.

Alink, A., Krugliak, A., Walther, A., & Kriegeskorte, N. (2013). fMRI orientation decoding in V1 does not require global maps or globally coherent orientation stimuli. *Frontiers in psychology, 4*.

Allport, A. (1993). Attention and control: Have we been asking the wrong questions? A critical review of twenty-five years. *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, 183-218.

Alvarez, I., de Haas, B., Clark, C. A., Rees, G., & Schwarzkopf, D. S. (2015). Comparing different stimulus configurations for population receptive field mapping in human fMRI. *Frontiers in human neuroscience, 9*.

Bahrami, B., Lavie, N., & Rees, G. (2007). Attentional load modulates responses of human primary visual cortex to invisible stimuli. *Current Biology, 17*(6), 509-513.

Bartels, A., Logothetis, N. K., & Moutoussis, K. (2008). fMRI and its interpretations: an illustration on directional selectivity in area V5/MT. *Trends in neurosciences, 31*(9), 444-453.

Berg, T. L., Berg, A. C., & Shih, J. (2010, September). Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision* (pp. 663-676). Springer Berlin Heidelberg.

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems* (pp. 2546-2554).

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *The journal of neuroscience*, 16(13), 4207-4221.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4), 324-345.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, 10, 433-436.

Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon press.

Brown, I. D. (2005). Review of the 'Looked but failed to see' accident causation factor (No. 60).

Campbell, F. W., Kulikowski, J. J., & Levinson, J. (1966). The effect of orientation on the visual resolution of gratings. *The Journal of physiology*, 187(2), 427.

Carmel, D., Saker, P., Rees, G., & Lavie, N. (2007). Perceptual load modulates conscious flicker perception. *Journal of Vision*, 7(14), 14-14.

Cartwright-Finch, U., & Lavie, N. (2007). The role of perceptual load in inattentive blindness. *Cognition*, 102(3), 321-340.

Chen, X., Bennett, P. N., Collins-Thompson, K., & Horvitz, E. (2013, February). Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 193-202). ACM.

Clevert, D. A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). Mathematical psychology: an elementary introduction.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

Dalal, N., & Triggs, B. (2004, June). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886-893). IEEE.

Dhar, S., Ordonez, V., & Berg, T. L. (2011, June). High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 1657-1664). IEEE.

de Fockert, J. W., Rees, G., Frith, C. D., & Lavie, N. (2001). The role of working memory in visual selective attention. *Science*, 291(5509), 1803-1806.

de Haas, B., Schwarzkopf, D. S., Anderson, E. J., & Rees, G. (2014). Perceptual load affects spatial tuning of neuronal populations in human early visual cortex. *Current Biology*, 24(2), R66-R67.

Dayan, P., & Solomon, J. A. (2010). Selective Bayes: Attentional load and crowding. *Vision research*, 50(22), 2248-2260.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1), 193-222.

Deutsch, J. A., & Deutsch, D. (1963). Attention: some theoretical considerations. *Psychological review*, 70(1), 80.

Ditterich, J., Mazurek, M. E., & Shadlen, M. N. (2003). Microstimulation of visual cortex affects the speed of perceptual decisions. *Nature neuroscience*, 6(8), 891-898.

Donahue, J., & Grauman, K. (2011, November). Annotator rationales for visual recognition. In *2011 International Conference on Computer Vision* (pp. 1395-1402). IEEE.

Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *Neuroimage*, 39(2), 647-660.

Du Tran, I., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2014). C3D: generic features for video analysis. *CoRR*, [abs/1412.0767](https://arxiv.org/abs/1412.0767), 2, 7.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub..

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598-601.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, 16(1), 143-149.

Ekstrom, A. (2010). How and when the fMRI BOLD signal relates to underlying neural activity: the danger in dissociation. *Brain research reviews*, 62(2), 233-244.

Estes, W. K. (2014). Mathematical models in psychology. *A Handbook for Data Analysis in the Behavioral Sciences: Volume 1: Methodological Issues Volume 2: Statistical Issues*, 3.

Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. (2009, June). Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (pp. 1778-1785). IEEE.

Farneback, G. (2003, June). Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis* (pp. 363-370). Springer Berlin Heidelberg.

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. *Science*, 322(5903), 970-973.

Forster, S., & Lavie, N. (2008). Failures to ignore entirely irrelevant distractors: the role of load. *Journal of Experimental Psychology: Applied*, 14(1), 73.

Forster, S., & Lavie, N. (2014). Distracted by your mind? Individual differences in distractibility predict mind wandering. *Journal of experimental psychology: learning, memory, and cognition*, 40(1), 251.

Frackowiak, R. S. (2004). *Human brain function*. K. J. Friston, C. D. Frith, R. J. Dolan, C. J. Price, S. Zeki, J. T. Ashburner, & W. D. Penny (Eds.). Academic press.

Friston, K. J., Holmes, A. P., Poline, J. B., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S., & Turner, R. (1995). Analysis of fMRI time-series revisited. *Neuroimage*, 2(1), 45-53.

Furmanski, C. S., & Engel, S. A. (2000). An oblique effect in human primary visual cortex. *Nature neuroscience*, 3(6), 535-536.

Gatti, S. V., & Egeth, H. E. (1978). Failure of spatial selectivity in vision. *Bulletin of the Psychonomic Society*, 11(3), 181-184.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. 1966. *New York*, 888, 889.

Green, M. (1981). Psychophysical relationships among mechanisms sensitive to pattern, motion and flicker. *Vision Research*, 21(7), 971-983.

Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., & Van Gool, L. (2013). The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1633-1640).

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7), 523-534.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1026-1034).

Heeger, D. J., & Ress, D. (2002). What does fMRI tell us about neuronal activity?. *Nature Reviews Neuroscience*, 3(2), 142-151.

Herbrich, R., Minka, T., & Graepel, T. (2006). Trueskill™: A Bayesian skill rating system. In *Advances in neural information processing systems* (pp. 569-576).

Hubel, D. H., & Wiesel, T. N. (1960). Receptive fields of optic nerve fibres in the spider monkey. *The Journal of physiology*, 154(3), 572-580.

Wiesel, T. N., & Hubel, D. H. (1963). Single-cell responses in striate cortex of kittens deprived of vision in one eye. *J Neurophysiol*, 26(6), 1003-1017.

Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology (London)*, 195, 215–243.

Hyder, F., Rothman, D. L., Mason, G. F., Rangarajan, A., Behar, K. L., & Shulman, R. G. (1997). Oxidative glucose metabolism in rat brain during single

forepaw stimulation: a spatially localized ^1H [^{13}C] nuclear magnetic resonance study. *Journal of Cerebral Blood Flow & Metabolism*, 17(10), 1040-1047.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11), 1254-1259.

Jiang, Y. G., Wang, Y., Feng, R., Xue, X., Zheng, Y., & Yang, H. (2013, June). Understanding and Predicting Interestingness of Videos. In *AAAI*.

Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4), 345-383.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5), 679-685.

Kang, K., Shapley, R. M., & Sompolinsky, H. (2004). Information tuning of populations of neurons in primary visual cortex. *The Journal of neuroscience*, 24(15), 3726-3735.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience*, 17(11), 4302-4311.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, 22(4), 751-761.

Kiapour, M. H., Yamaguchi, K., Berg, A. C., & Berg, T. L. (2014, September). Hipster wars: Discovering elements of fashion styles. In *European conference on computer vision* (pp. 472-488). Springer International Publishing.

Klauer, S. G., Neale, V. L., Dingus, T. A., Ramsey, D., & Sudweeks, J. (2005, September). Driver inattention: A contributing factor to crashes and near-crashes. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, No. 22, pp. 1922-1926). SAGE Publications.

Kok, P., Jehee, J. F., & de Lange, F. P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, 75(2), 265-270.

Kovashka, A., Parikh, D., & Grauman, K. (2012, June). Whittlesearch: Image search with relative attribute feedback. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 2973-2980). IEEE.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008, June). Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1-8). IEEE.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human perception and performance*, 21(3), 451.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in cognitive sciences*, 9(2), 75-82.

Lavie, N. (2010). Attention, distraction, and cognitive control under load. *Current Directions in Psychological Science*, 19(3), 143-148

Lavie, N., & Cox, S. (1997). On the efficiency of visual selective attention: Efficient visual search leads to inefficient distractor rejection. *Psychological Science*, 8(5), 395-396.

Lavie, N., & De Fockert, J. W. (2003). Contrasting effects of sensory limits and capacity limits in visual selective attention. *Perception & Psychophysics*, 65(2), 202-212.

Lavie, N., Lin, Z., Zokaei, N., & Thoma, V. (2009). The role of perceptual load in object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 35(5), 1346.

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics*, 56(2), 183-197.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10), 1995.

- Lee, J., & Maunsell, J. H. (2010). Attentional modulation of MT neurons with single or multiple stimuli in their receptive fields. *The Journal of Neuroscience*, 30(8), 3058-3066.
- Lennie, P. (2003). The cost of cortical computation. *Current biology*, 13(6), 493-497.
- Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the BOLD signal. *Annu. Rev. Physiol.*, 66, 735-769.
- Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on* (Vol. 2, pp. 1150-1157). IEEE.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91-110.
- Luce, R. D. (1959). On the possible psychophysical laws. *Psychological review*, 66(2), 81.
- Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of neurophysiology*, 77(1), 24-42.
- Macdonald, J. S., & Lavie, N. (2008). Load induced blindness. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5), 1078.
- Mack, A., & Rock, I. (1998). *Inattention blindness* (Vol. 33). Cambridge, MA: MIT press.

Mangun, G. R. (1995). Neural mechanisms of visual selective attention. *Psychophysiology*, 32(1), 4-18.

Marciano, H., & Yeshurun, Y. (2012). Perceptual load in central and peripheral regions and its effects on driving performance: Advertising billboards. *Work*, 41(Supplement 1), 3181-3188.

Marciano, H., & Yeshurun, Y. (2015). Perceptual load in different regions of the visual scene and its relevance for driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(4), 701-716.

Marquart, G., Cabrall, C., & de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, 3, 2854-2861.

Marszalek, M., Laptev, I., & Schmid, C. (2009, June). Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (pp. 2929-2936). IEEE.

Martinez-Trujillo, J. C., & Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology*, 14(9), 744-751.

Mazur, L. M., Mosaly, P. R., Jackson, M., Chang, S. X., Burkhardt, K. D., Adams, R. D., ... & Marks, L. B. (2012). Quantitative assessment of workload and stressors in clinical radiation oncology. *International Journal of Radiation Oncology* Biology* Physics*, 83(5), e571-e576.

- McAdams, C. J., & Maunsell, J. H. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology*, 83(3), 1751-1755.
- McMahon, M. J., & Macleod, D. I. (2003). The origin of the oblique effect examined with pattern adaptation and masking. *Journal of Vision*, 3(3), 4-4.
- Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage*, 53(1), 103-118.
- Monti, M. M. (2011). Statistical analysis of fMRI time-series: a critical review of the GLM approach. *Frontiers in human neuroscience*, 5(28).
- Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Frontiers in cognitive neuroscience*, 229, 342-345.
- Murphy, G., & Greene, C. M. (2016). Perceptual Load Induces Inattentional Blindness in Drivers. *Applied Cognitive Psychology*, 30(3), 479-483.
- Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive psychology*, 7(1), 44-64.
- Nowak, S., & R ger, S. (2010, March). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval* (pp. 557-566). ACM.

Obleser, J., Leaver, A., VanMeter, J., & Rauschecker, J. P. (2010). Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. *Frontiers in psychology, 1*, 232.

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature neuroscience, 5*(11), 1203-1209.

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences, 87*(24), 9868-9872.

Panzeri, S., Magri, C., & Logothetis, N. K. (2008). On the use of information theory for the analysis of the relationship between neural and imaging signals. *Magnetic resonance imaging, 26*(7), 1015-1025.

Parks, N. A., Hilimire, M. R., & Corballis, P. M. (2011). Steady-state signatures of visual perceptual load, multimodal distractor filtering, and neural competition. *Journal of Cognitive Neuroscience, 23*(5), 1113-1124.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision, 10*(4), 437-442.

Phelps, A. S., Naeger, D. M., Courtier, J. L., Lambert, J. W., Marcovici, P. A., Villanueva-Meyer, J. E., & MacKenzie, J. D. (2015). Pairwise comparison versus Likert scale for biomedical image assessment. *American Journal of Roentgenology, 204*(1), 8-14.

Pinsk, M. A., Doniger, G. M., & Kastner, S. (2004). Push-pull mechanism of selective attention in human extrastriate cortex. *Journal of Neurophysiology*, 92(1), 622-629.

Raveh, D., & Lavie, N. (2015). Load-induced inattentional deafness. *Attention, Perception, & Psychophysics*, 77(2), 483-492.

Rees, G., Friston, K., & Koch, C. (2000). A direct quantitative relationship between the functional properties of human and macaque V5. *Nature neuroscience*, 3(7), 716-723.

Rees, G., Frith, C. D., & Lavie, N. (1997). Modulating irrelevant motion perception by varying attentional load in an unrelated task. *Science*, 278(5343), 1616-1619.

Roper, Z. J., Cosman, J. D., & Vecera, S. P. (2013). Perceptual load corresponds with factors known to influence visual search. *Journal of experimental psychology: human perception and performance*, 39(5), 1340.

Sereno, M. I., Dale, A. M., Reppas, J. B., & Kwong, K. K. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212), 889.

Saproo, S., & Serences, J. T. (2010). Spatial attention improves the quality of population codes in human visual cortex. *Journal of neurophysiology*, 104(2), 885-895.

Scalf, P. E., Torralbo, A., Tapia, E., & Beck, D. M. (2013). Competition explains limited attention and perceptual resources: implications for perceptual load and dilution theories. *Frontiers in psychology*, 4, 243.

Schwartz, S., Vuilleumier, P., Hutton, C., Maravita, A., Dolan, R. J., & Driver, J. (2005). Attentional load and sensory competition in human vision: modulation of fMRI responses by load at fixation during task-irrelevant stimulation in the peripheral visual field. *Cerebral cortex*, 15(6), 770-786.

Serences, J. T., Saproo, S., Scolari, M., Ho, T., & Muftuler, L. T. (2009). Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. *Neuroimage*, 44(1), 223-231.

Seriès, P., Latham, P. E., & Pouget, A. (2004). Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nature neuroscience*, 7(10), 1129-1135.

Scholl, B., Tan, A. Y., Corey, J., & Priebe, N. J. (2013). Emergence of orientation selectivity in the mammalian visual pathway. *The Journal of Neuroscience*, 33(26), 10616-10624.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of neurophysiology*, 86(4), 1916-1936.

Shotton, J., Johnson, M., & Cipolla, R. (2008, June). Semantic texton forests for image categorization and segmentation. In *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1-8). IEEE.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentive blindness for dynamic events. *Perception*, 28(9), 1059-1074.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Snyder, C. R. (1972). Selection, inspection, and naming in visual search. *Journal of Experimental Psychology*, 92(3), 428.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological monographs: General and applied*, 74(11), 1.

Sprague, T. C., Saproo, S., & Serences, J. T. (2015). Visual attention mitigates information loss in small-and large-scale neural codes. *Trends in cognitive sciences*, 19(4), 215-226.

Stolte, M., Bahrami, B., & Lavie, N. (2014). High perceptual load leads to both reduced gain and broader orientation tuning. *Journal of vision*, 14(3), 9-9.

Sunny, M. M., & von Mühlhausen, A. (2014). The role of flicker and abrupt displacement in attention capture by motion onsets. *Attention, Perception, & Psychophysics*, 76(2), 508-518.

Tang, S., & Huang, L. L. (2013, November). Traffic sign recognition using complementary features. In *2013 2nd IAPR Asian Conference on Pattern Recognition* (pp. 210-214). IEEE.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, 34(4), 273.

- Torralba, A., Oliva, A., Castelhana, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4), 766.
- Torralba, A., & Beck, D. M. (2008). Perceptual-load-induced selection as a result of local competitive interactions in visual cortex. *Psychological science*, 19(10), 1045-1050.
- Treisman, A. M., & Riley, J. G. (1969). Is selective attention selective perception or selective response? A further test. *Journal of Experimental Psychology*, 79(1p1), 27.
- Treue, S., & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736), 575-579.
- Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nature neuroscience*, 8(8), 1096-1101.
- Vanderbei, R. J. (1999). LOQO: An interior point code for quadratic programming. *Optimization methods and software*, 11(1-4), 451-484.
- Vincent, E., & Laganière, R. (2001, June). Detecting planar homographies in an image pair. In *Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis* (pp. 182-187).
- Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2011, June). Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 3169-3176). IEEE.

Wang, Y., & Mori, G. (2009). Human action recognition by semilattent topic models. *IEEE transactions on pattern analysis and machine intelligence*, 31(10), 1762-1774.

Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3551-3558).

Wang, H., Oneata, D., Verbeek, J., & Schmid, C. (2015). A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, 1-20.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & psychophysics*, 33(2), 113-120.

Yamaguchi, K., Okatani, T., Sudo, K., Murasaki, K., & Taniguchi, Y. (2015). Mix and match: joint model for clothing and attribute recognition. In *British Machine Vision Conf.*

Yi, D. J., Woodman, G. F., Widders, D., Marois, R., & Chun, M. M. (2004). Neural fate of ignored stimuli: dissociable effects of perceptual and working memory load. *Nature neuroscience*, 7(9), 992-996.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., ... & Torr, P. H. (2015). Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1529-1537).