THE UNIVERSITY *of* EDINBURGH

Modelling Eye Movements and Visual Attention in Synchronous Visual and Linguistic Processing

Michal Dziemianko

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2013

Abstract

This thesis focuses on modelling visual attention in tasks in which vision interacts with language and other sources of contextual information. The work is based on insights provided by experimental studies in visual cognition and psycholinguistics, particularly cross-modal processing.

We present a series of models of eye-movements in situated language comprehension capable of generating human-like scan-paths. Moreover we investigate the existence of high level structure of the scan-paths and applicability of tools used in Natural Language Processing in the analysis of this structure.

We show that scan paths carry interesting information that is currently neglected in both experimental and modelling studies. This information, studied at a level beyond simple statistical measures such as proportion of looks, can be used to extract knowledge of more complicated patterns of behaviour, and to build models capable of simulating human behaviour in the presence of linguistic material.

We also revisit classical model saliency and its extensions, in particular the Contextual Guidance Model of Torralba et al. (2006), and extend it with memory of target positions in visual search. We show that models of contextual guidance should contain components responsible for short term learning and memorisation. We also investigate the applicability of this type of model to prediction of human behaviour in tasks with incremental stimuli as in situated language comprehension.

Finally we investigate the issue of *objectness* and object saliency, including their effects on eye-movements and human responses to experimental tasks. In a simple experiment we show that when using an object-based notion of saliency it is possible to predict fixation locations better than using pixel-based saliency as formulated by Itti et al. (1998). In addition we show that object based saliency fits into current theories such as cognitive relevance and can be used to build unified models of cross-referential visual and linguistic processing.

This thesis forms a foundation towards a more detailed study of scan-paths within an object-based framework such as Cognitive Relevance Framework (Henderson et al., 2007, 2009) by providing models capable of explaining human behaviour, and the delivery of tools and methodologies to predict which objects would be attended to during synchronous visual and linguistic processing.

Acknowledgements

My first thanks go to my supervisors Frank Keller and Hiroshi Shimodaira, for their numerous suggestions, inspiring ideas, and interesting discussions. I am grateful for a chance to work with these world class scientists.

I would also like to thank Antje Nuthmann, Robin Hill and Ben Tatler for their constructive criticism and feedback provided on various occasions.

Special thanks go to my colleagues Moreno Coco, Alasdair Clarke and John Pate for sharing relevant data, source code and their help with solving various problems.

I would also like to mention Brent Kevit-Kylar, Trevor Fountain, Ben Allison, Ali Eslami, Desmond Elliott and many more who have given me an input throughout all these years.

A final thanks go to my wife Klaudia and my family for their love and support.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Michal Dziemianko*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Our everyday lives require processing a vast amount of information coming from different senses at the same time. Such cross-modal synchronous processing is necessary to perform the majority of daily tasks. For example, when pouring a coffee into a cup our vision provides information that is necessary to coordinate motor actions. Similarly, while driving a car, visual information systems share and coordinate with systems controlling our motor actions in order to steer the car, governing our route planning, and many other processes.

Even though almost all our daily tasks require cross-modal processing, our knowledge of this process is incomplete. Due to the topic being very broad and complex, we will only focus on interactions between two specific processing paths: visual and linguistic.

Experimental studies of synchronous linguistic processing traditionally focused on linguistic context, and phenomena like influence of discourse on disambiguation (e.g. Altmann and Steedman, 1988; Hare et al., 2003; Binder et al., 2001). However, people also refer to objects present in the surrounding world, as well as events and relations between them (Gleitman, 1990). Most of the information about these objects comes from vision, and the first studies of the influence of visual context date back to at least mid-1970 (Cooper, 1974). Nonetheless, at the time, eye-tracking technology was inaccurate and expensive. Therefore it was not widely available and accepted in linguistics, heavily limiting the number of experimental studies.

More recently, the Visual World Paradigm (VWP) (Tanenhaus et al., 1995) is widely exploited in numerous psycholinguistic studies. With its help, clear links between visual and linguistic processing have been shown (see e.g. Altmann and Kamide, 1999; Altmann and Mirkovic, 2009; Crocker et al., 2010; Knoeferle and Crocker, 2006;

Spivey-Knowlton et al., 2002; Snedeker and Trueswell, 2003). Moreover linguistic processing triggers characteristic eye-movement responses, with fixation patterns reflecting, for instance, syntactic disambiguation (e.g. Tanenhaus et al., 1995; Coco, 2011) or semantic anticipation (e.g. Sedivy et al., 1999).

Computational models of synchronous processing exist, however they usually focus on linguistic phenomena or are extensions of language processing models used to account for some visual factors. For example, the FUSE model of Roy and Mukherjee (2005) uses visual context in order to improve speech recognition. Using a mechanism conceptually similar to visual attention, it exploits contents of the visual scene and provides prior probabilities based on a language model to constrain possible interpretations of an acoustic signal.

Similarly Mayberry et al. (2005) describes a model of anticipation capable of integrating visual information into prediction of upcoming referents (i.e. objects mentioned in the sentence).

Recently more interest was put towards investigation of speech in a natural or close to natural setting. A series of papers (e.g. Qu and Chai, 2009, 2008; Prasov and Chai, 2008; Prasov et al., 2007) investigate the role of visual attention in multi-modal conversation interfaces. They show that eye-tracking can easily improve speech recognition, and allows automatic lexicon acquisition.

Complex and accurate models of visual processing itself are however quite rare, especially those capable of predicting eye movements. The most prominent example is the work of Kukona and Tabor (2011), who present an analysis of human behaviour in a simple VWP task along with a computational model. However the presented connectionist model is very limited - it only recognises 5 hand picked words and is based on a hand-wired neural network. Nevertheless it is able to predict the proportion of fixations falling onto referred objects, their direct competitors, and distractors.

At the same time it was shown in visual cognition, that visual attention is driven by top-down, contextual processes (see e.g. Schyns and Oliva, 1994; Oliva and Torralba, 2007), with several modelling attempts exist including Torralba et al. (2006); Chanceaux et al. (2008).

It has also been shown that visual attention is driven in an object based top-down fashion (see e.g. Henderson et al., 1999; Henderson, 2003; Nuthmann and Henderson, 2010; Zelinsky and Schmidt, 2009; Findlay and Gilchrist, 2001; Einhauser et al., 2008). This has led to the emergence of new views such as the Cognitive Relevance hypothesis (Henderson et al., 2007, 2009). No complete models compatible with rel-

evance hypothesis exists. The most advanced is work of Spain and Perona (2011) which attempts to predict which objects are mentioned during a naming experiment. In addition several models of object-based attentional selection has been proposed in Robotics and Computer Vision such as Schauerte et al. (2012); Yu et al. (2010). Explanations for off-object fixations has been also proposed such as population averaging modelled by *Target Acquisition Model*(TAM) (Zelinsky, 2008), or oculomotor system errors (Nuthmann and Henderson, 2010).

It is also known, that eye movement patterns depend on the task performed (Yarbus, 1967; Castelhano et al., 2009; Schmidt and Zielinsky, 2009). However, the studies and modelling in visual cognition field are usually based on tasks simpler than synchronous processing, such as visual search, and rarely involve language.

In addition to research in psycholinguistics and visual cognition, studies in other areas have yielded results suggesting that analysis and modelling of visual attention and eye movements in synchronous processing is not only possible, but highly desirable. For example Leek et al. (2012) present an eye-tracking study of a grasping task, where eye movements are found to be predictive of how the objects is going to be grasped. Simola et al. (2008) present Hidden Markov Model (HMM) based classifier capable of discriminating between various tasks performed by people over the course of eye-tracking experiment.

1.1 Central claims

Interest in situated language processing resulted in multiple experimental and modelling studies of linguistic and visual processing. However models of visual attention integrating high-level contextual guidance and capable of accurate prediction of eye-movements are relatively rare. Such models of visual attention capable of dealing with linguistic context are almost non existent.

This thesis aims to build a foundation towards a more detailed study of visual attention by providing models capable of explaining human behaviour such as scan-paths and to predict which objects are attended in cross-modal referential processing.

This thesis puts forward three main claims. The first claim is that scan-paths carry interesting information. Moreover they are consistent across people enough to allow recovery and use of this information e.g. to build computational models.

The second claim is that linguistic input, particularly its interpretation at any given moment, is crucial to study and synthesise human scan-paths during situated language

processing.

The third claim is that an object's relevance is the main factor governing attentional selection. However bottom-up saliency should not be discarded, but rather treated as a feature of an object instead of as an area or separate pixels. This is especially important for synchronous visual and linguistic processing due to the referentiality of a language - the cooperation of language and vision is based on objects, rather than arbitrary patches.

In this thesis we do not attempt to build one coherent, cognitively plausible model of synchronous processing, but rather provide a series of models addressing problems that are underestimated in similar studies. We put particular focus on prediction of eye-movements, including the sequential order of fixations, and object-based architecture, that is compatible with recent experimental evidence of top-down attentional selection.

1.2 Motivation

Computational modelling is an important methodology. It helps to formalize concepts, observe effects of assumptions, parameters and understand their implications. Implementation of models results in various questions or uncertainties, that lead to well motivated experimental studies. Moreover modelling helps to integrate multiple partial frameworks and theories into one compatible set of tools.

A complete, working model would allow the investigate of which factors account for certain behavioural patterns and, in turn, achieve a better understanding of how visual attention works in the presence of linguistic material. Predictions on new, unseen materials help, on the other hand, with getting an intuition on what human behaviour to expect during various experimental conditions.

Accurate models of visual attention in situated language comprehension can be potentially applied to dialogue systems. Comparing real human behaviour with expectations based on model predictions, can be used as a basis of a system testing the level of understanding achieved between speaking parties. In case of detected misunderstanding or other unusual behavioural pattern that might indicate a potential problem, the system can rephrase and repeat relevant pieces of information. Similarly, through studying the behaviour of a speaker, it is potentially possible to achieve a better understanding of human input in situations relying on visual context such as navigation.

In e-learning, predictions might be used to refine visual and linguistic material in order to ensure that important information is presented in a form that allows its easier

and more efficient comprehension.

1.3 Overview of the thesis and contributions

Chapter 2, presents tools and methods used as a basis for development of models presented in this thesis.

Datasets used in development and evaluation of created models including associated issues such as data encoding schemes, are presented in chapter 3.

Chapter 4 investigates contextual guidance as formulated and modelled by Torralba et al. (2006). It mainly contains the derivation of *Memory Modulated Saliency* (MMS, see Dziemianko et al., 2011b) - an extension of *Contextual Guidance Model* (Torralba et al., 2006, CGM) to account for short term online learning of the arrangement of the objects in the scenes during the visual search task.

A possibility of extending this framework to handle incremental stimuli such as that used in the Visual World Paradigm experiments is also investigated in this chapter. The initial results show, that the model is capable of predicting fixation locations with performance exceeding all baseline levels.

Chapter 5 continues work towards a model of eye movements in situated language comprehension. We derive a sequence of models that are able to generate human-like eye-movements. The experimental evaluation shows, that it is possible to learn and synthesise scan-paths. Moreover we show the importance of linguistic input and its interpretation.

Chapter 6 investigates the hypothesis of higher level structure being present in scan-paths. It shows that existing techniques - namely shallow parsing - are capable of learning this structure. It also shows that, this structure might be successfully used to improve scan-path synthesis, as well as perform other tasks such as a classification of the listener.

In addition, section 6.4 presents applications of shallow parsing to the study of similarity between scan-paths. A proof-of-concept method is implemented that exhibits promising performance, despite its simplicity.

Finally the chapter 7 investigates an issue of *objectness* and an object's saliency and the effect this has on eye-movements and human responses to high-level experimental tasks. It is shown, that an object based notion of saliency is a better predictor of fixation locations, than pixel-based saliency as formulated by Itti et al. (1998).

Moreover the chapter shows how object-based saliency can be integrated with mod-

els presented in chapters 5 and 6 in order to create models integrating visual and linguistic processing paths for eye-movement prediction.

1.4 Collaborations and Publications

The experiments presented in chapter 4 were partially published in Dziemianko et al. (2011b). An extended work benefited from comments from reviewers of *Visual Cognition* and has been published in Dziemianko and Keller (2013).

Models presented in chapter 5 were presented as abstract and discussed with audience of CUNY-2012 conference.

Implementation of scan-path shallow parsers presented in chapter 6 is based on code provided by John K. Pate.

Work described in chapter 7 was partially done in collaboration with Alasdair Clarke. Results discussed in section 7.5 were published in Dziemianko et al. (2013) and an extended abstract was presented at *Predicting Perceptions* 2012 conference.

Chapter 2

Methodology and Tools

2.1 Introduction

In this chapter we describe methods used in the rest of this thesis for analysis and modelling of experimental material. In section 2.2.3 we discuss two common representations of eye movements - fixation densities and scan-paths, and essential differences between them.

As this thesis focuses mainly on scan-paths, we discuss the impact of this representation on analysis and modelling of experimental data. Section 6.4 discusses the problem of scan-path comparison with emphasis on the necessity for object based processing.

We also briefly discuss statistical tools and methodologies used to analyse the available data and modelling results.

Finally we present a discussion of the importance of computational modelling in both: context of language and visual processing, with emphasis on attention.

2.2 Eye-movements and eye-tracking

2.2.1 Eye movements as part of attentional mechanism

Understanding a visual world requires a considerable amount of real-time processing of rich and highly dynamic information. Even though considerable parts of our brain are devoted to vision and other sensory information, we are not able to process all the possible information fast enough for us to respond to all the perceived objects at once. Therefore most of the available visual field is processed coarsely, with an additional

mechanism for selecting objects of interest for further high level processing (see e.g. Posner, 1978; LaBerge, 1983; Johnston and Dark, 1986).

Such a mechanism is closely related to eye-movements. Our eye can only deliver a sharp, full resolution image of the surrounding world within a small area, the *fovea centralis*. Even though the fovea delivers images of less than 2 degrees of visual field, it requires over 50% of the visual cortex to process this information (Krantz, 2012). Eye movements called saccades allow us to access larger portions of the visual field by capturing 3-4 images of different areas per second. These areas are than integrated by the brain, giving an impression of sharp, and high resolution visual capabilities within the whole visual field.

Visual attention allows us to select objects or areas, that would benefit from being processed with the foveal region and further, high level processing such as object recognition. It is important to note, that visual attention is not strictly confined to to region processed by the fovea, but instead it is still possible to attend to regions within a few degrees from its centre (Posner, 1978). It is however very important to realise, that attentional shifts and eye movements are closely related, and occur together, or follow each other in close succession (see e.g. Anderson et al., 1995; Posner, 1980).

Thus, eye motion can be treated as an indication of visual attention and, by extrapolation, of cognitive processes.

2.2.2 Eye-tracking

Eye tracking enables us to study the allocation of attention on presented stimuli. Both spatial and temporal aspects can be captured, with *saccade* and *fixation* being of the main interest.

The saccade is an eye movement between two consecutive locations. It is usually measured in terms of distance (e.g. degrees of visual field), with the duration being of less importance as it is directly correlated with the distance.

The fixation is a time, when eye gaze stays still[1] focusing on a certain location for a certain period of time. The fixation durations are equally as important as the spatial location, and are often believed to be associated with *cognitive processing* by indicating *processing load* (Rayner, 1998). The exact fixation durations vary considerably, with the average being 200-300ms depending, among others, on the performed task (Castelhano et al., 2009).

[1] With exception of small oscillations that occur all the time.

Eye tracking is a procedure that allows us to capture the direction of an eye gaze and its position relative to the head and presented stimuli (e.g. image displayed on a screen). The eye trackers do not output information about fixations and saccadic movement directly. Instead they provide information of spatial position of an eye gaze with a constant sampling rate. The eye tracker used to collect most of the datasets used in this study - EyeLink II - works with a temporal resolution of 500 samples per second. The detection of saccades is performed by analysis of eye displacement. The off-the-shelf configuration of EyeLink classifies all motion with a rate greater than a specified threshold as saccade. Eyelink manual (SR Research, 2002) recommends using a threshold of 22 degrees of visual field per second. Similarly periods between saccades are considered to be fixations.

Such a procedure is an oversimplification of real eye gaze motion dynamics (e.g. see Otero-Millan et al., 2008, for discussion of microsaccades), however it is acceptable for studies of visual attention.

2.2.3 Representing eye movements: fixation densities and scan-paths

One of the most important aspects of eye tracking analysis is the selection of appropriate representations of the saccades and fixations captured during the course of the experimental trial. We will further refer to the sequence of such alternating saccades and fixations as *scan-paths*.

Preserving the whole dynamics of the scan-path - i.e. durations and spatial coordinates of fixations and saccades involves certain problems, therefore simplified representations and means of analysis are often used, discarding all the information outside the scope of interest of the particular study.

Commonly used methods involve analysing distributional characteristics of properties such as saccade amplitude, fixation durations, proportion of looks falling onto certain region etc. Fixation densities, commonly represented as *heat maps* are also among popular methods of analysis and visualisation of the behaviour.

Even though these methods allow us to study certain phenomena such as the association of looks at certain image regions with certain events, they suffer from a very serious problem: the temporal information i.e. ordering of fixations and dynamics of behaviour over the time are lost. Therefore these representations are applicable only if we are interested in statistical analysis of the behaviour of experimental subjects.

Full representations of scan-paths - preserving both spatial and temporal information - are much richer, however less popular due to numerous reasons. Firstly, since the early work of Yarbus (1967), fixation densities have been found to be relatively consistent among participants performing the same task, while fixations order is traditionally believed to be heavily variable. Secondly there is no established, widely recognised and accepted procedure for comparing two scan-paths. Preserving the order of fixations enables us to study and discover more subtle patterns of behaviour e.g. competition of objects for attention.

Figure 2.1 presents a visualisation of eye-tracking data represented as a heat-map and as full scan paths.

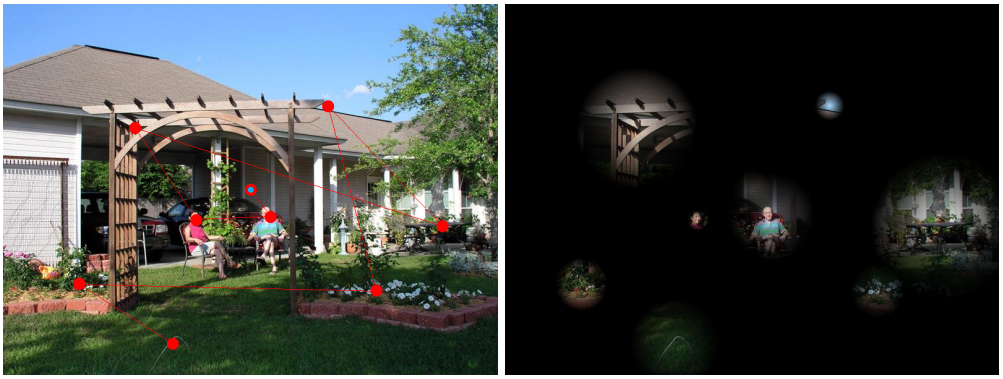


Figure 2.1: A visualization of data represented as scanpath (left) and as fixations heatmap (right). The major difference - information about ordering of the fixations in scan-path can be noticed immediately. Both representations are capable of representing fixation duration - scan-paths directly, and heat maps indirectly e.g. by amplitude or size of the hot spot.

In addition, heat maps allow an easy representation of collective behaviour by layering several maps together. Such compound heat maps provide valuable information about general behaviour and preferred fixation locations across all the experimental subjects. Scan paths, on the other hand, can not be combined in such an easy way without loosing associated temporal information.

2.3 Modelling

2.3.1 Hidden Markov Models

Hidden Markov Models (HMMs, for complete tutorial see Rabiner, 1989) form a very important class of models being widely applied in statistics, pattern recognition, speech processing and many other fields. HMMs are simplest form of dynamic Bayesian Networks. and is generalization of a mixture model, where latent variables are not independent, but rather related to each other through *Markov process*.

2.3.1.1 Markov property and Markov Process

A key concept in the whole theory of HMMs is a *Markov process*. A Markov process is a stochastic process that satisfy the *Markov property*. The Markov property states, that future states of the process depend only upon the present state, and not on the whole sequence of states that preceded it. This is expressed by means of the conditional probability:

$$P(s_T|s_{T-1}, s_{t-2}, ..., s_1, s_0) \quad = \quad P(s_T|s_{T-1}) \tag{2.1}$$

where s_t is state of the process at discrete time t.

2.3.1.2 Alignment and tagging with Hidden Markov Models

A statistical model that has a Markov property is called a Markov model. Regular Markov models have the state sequence observable, thus the only parameters are the transition probabilities. In HMMs the states are not directly visible, instead only the output, dependent on these states, is visible. Each state is associated with a certain probability distribution over possible output outcomes (symbols), therefore sequences of output symbols provide some information about the state sequence.

Through this thesis HMMs are used to solve two main tasks: alignment and tagging. We will first explain how HMMs are applied to solve the alignment task.

Let each entity ω_i to be aligned (i.e. word from vocabulary $\Omega = \omega_1, \omega_2, ..., \omega_N$) with a sequence of observations $O = o_1, o_2, ..., o_{T-1}, o_T$. The alignment task can be then defined as finding $\hat{\omega}$ such that:

$$\hat{\omega} \quad = \quad \arg\max_{\omega_i} P(\omega_i|O) \tag{2.2}$$

In practice probability $P(\omega|O)$ is not computable directly, however it can be obtained using Bayes' Rule:

$$P(\omega_i|O) = \frac{P(O|\omega_i)P(\omega_i)}{P(O)} \tag{2.3}$$

Effectively stating that for given prior probabilities $P(\omega)$, the most probable word ω depends exclusively on the likelihood $P(O|\omega)$.

Computation of conditional probability $P(O|\omega) = P(o_1, o_2, ..., o_T|\omega)$ might be not tractable, however assuming that sequence $O = o_1, o_2, ..., o_T$ is generated by Markov process reduces the problem to estimation of Markov Model parameters.

A Markov Model can be seen as a Finite State Machine (FSM) where at each time unit t the transition from state i to state j is occurring with probability a_{ij}, and an observation vector o_t is generated from emission probability distribution $b_j(o_t)$. Figure 2.2 presents graphical representation of such FSM.

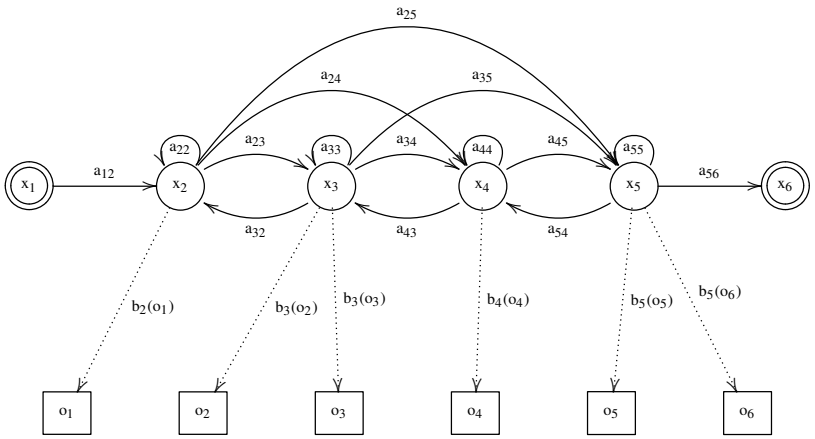


Figure 2.2: An example of HMM represented as Finite State Machine. $O = (o_1, ..., o_T)$ is an observable stream generated by an underlying state sequence $X = (x_1, ..., x_T)$, a_{ij} are transition probabilities between states x_i and x_j, and $b_i(o_j)$ are emission probabilities of symbols o_j at states x_i.

The probability of the observed sequence O being generated by a Markov Model M by passing state sequence X can be expressed as a product of transition and emission probabilities. However, as already mentioned, in Hidden Markov Models the state sequence X is not directly visible, and the required likelihood must be computed by summing over all possible state sequences $X = x_1, ...x_T$:

$$P(O|M) = \sum_X a_{x_0 x_1} \prod_{t=1}^{T} b_{x_t}(o_t) a_{x_t x_{t+1}} \tag{2.4}$$

where x_0 and x_{T+1} are model entry and exit states.

In practical applications it is often possible, or even desirable to consider only the most probable state sequence:

$$P(O|M) \approx \max_X a_{x_0 x_1} \prod_{t=1}^{T} b_{x_t}(o_t) a_{x_t x_{t+1}} \tag{2.5}$$

The recognition task can be than easily solved by building a set of models M_i corresponding to our initial vocabulary ω_i by assuming:

$$P(O|w_i) = P(O|M_i) \tag{2.6}$$

and in turn:

$$\hat{\omega} = \arg\max_{\omega} P(\omega_i|O) = \arg\max_{M_i} P(M_i|O) \tag{2.7}$$

The remaining issue is efficient computation of the aforementioned probabilities and estimation of model parameters from training data.

The second of our tasks - tagging - can be formulated as an assignment of each observed symbol o_t to one of the classes (tags) $c_j(o_t)$ considering previously seen part of the observed sequence:

$$\hat{c}(o_t) = \arg\max_{c_j(o_t)} P(c_j(o_t)|o_1, o_2, ..., o_{t-1}) \tag{2.8}$$

or more generally finding the sequence of tags $C = c_1(o_1), c_2(o_2), ..., c_T(o_T)$ corresponding to observed sequence $O = o_1, o_2, ..., o_T$:

$$\hat{C} = \arg\max_{C} P(C|O) \tag{2.9}$$

We will reduce this problem to a recovery of the most probable sequence of HMM states X by assuming, that each state x_i of the FSM representing HMM corresponds to a specific tag c_i, hence:

$$\hat{C} = \arg\max_{C} P(C|O) \tag{2.10}$$

$$= \arg\max_{X} P(X|O) \tag{2.11}$$

$$= \arg\max_{X} a_{x_0 x_1} \prod_{t=1}^{T} b_{x_t}(o_t) a_{x_t x_{t+1}} \tag{2.12}$$

Alternatively tagging can be seen as a continuous recognition task i.e. observable stream O does not correspond to output from one model M, but rather the concatenated

outputs $O^1, O^2, ..., O^\tau$ corresponding to the sequence of models $M_1, M_2, ..., M^\tau$. The operation of HMMs is the same in this case as for alignment, however the calculation of observed sequence probability requires not only recovery of the most probable state transition, but also the most probable set of HMMs used to generate each part of this sequence. This task can be solved by an extension of Viterbi algorithm.

It is important to notice, that the first formulation is more suitable to cases where each output symbol should have a corresponding tag assigned. The latter is however more suitable in cases where subsequence of output symbols form larger groups, that should be tagged. In this work we use both formulations as appropriate.

Through this work we will assume that transition probabilities are discrete. However, the emission probabilities in continues density HMMs, are commonly represented by Mixture of Gaussians. Emission probabilities $b_j(o_t)$ can be therefore expressed as:

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(o_t; \mu_{jm}, \Sigma_{jm}) \tag{2.13}$$

where M is number of mixture components, c_{jm} is weight of m-th component, and $\mathcal{N}(*; \mu, \Sigma)$ is a multivariate Gaussian with mean μ and covariance Σ:

$$\mathcal{N}(o, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o-\mu)' \Sigma^{-1} (o-\mu)} \tag{2.14}$$

The estimation of the transition and emission probabilities (i.e. model parameters) is normally done with *Baum-Welch Re-Estimation* algorithm. The calculation of sequence probabilities given trained model is achieved using the Viterbi algorithm. The explanation and derivation of these algorithms is however out of the scope of this thesis. Detailed information about HMMs and the algorithms can be found, among others in Rabiner (1989); Young et al. (2006).

2.3.2 Markov Chain Monte Carlo (MCMC) methods

Markov Chain Monte Carlo methods are a class of algorithms used for sampling. They are based on idea of constructing Markov chains converging on the desired *equilibrium distribution*. Assuming a sufficiently long Markov chain, the state is considered as a sample from the desired distribution.

The key issue in MCMC methods is to determine how many steps are to be taken to converge on the desired distribution within acceptable time from any starting position. The approximation improves with the number of states, however it comes with a cost of longer execution time. However, such an approximation might be the only practical

way of estimating complex joint probability distributions, for which exact solutions might be computationally infeasible.

The simplest MCMC methods are based on the idea of generating samples using *random walk*. Through this work we will use the *Metropolis-Hastings* algorithm, and *Gibbs sampling*. Their main drawback - that they possibly need a long time to explore whole probability space is not an issue in our applications, therefore we do not require any more advanced methods.

2.3.2.1 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a method for sampling from a probability distribution for which it is difficult to obtain samples directly. The main advantage of the algorithm is, that it can draw a sample z from the probability distribution $p(z)$ provided only a function $\hat{p}(z)$ proportional to $p(z)$ with some constant Z_p which can be easily calculated.

In this method, samples are generated from a proposed distribution. During the sampling we keep track of the current state $z^{(\tau)}$ and the proposed distribution $q(z|z^{(\tau)})$ depends on the current state. As an effect of this, sequence of states $z^{(0)}, z^{(1)}, ..., z^{(\tau)}$ form a Markov chain.

Writing $p(z) = \frac{\hat{p}(z)}{Z_p}$, and assuming that function $\hat{p}(z)$ can easily be evaluated, we choose a proposed distribution $q(z)$ such that it is easy to draw samples directly from it. At each step of algorithm, a sample z^* is generated from the proposed distribution, and then accepted or rejected according to certain criterion.

If we assume that the proposed distribution is symmetric i.e. $q(z_a|z_b) = q(z_b|z_a)$ for all values of z_a and z_b the candidate sample z^* can be accepted with probability:

$$A(z*, z^{(\tau)}) \quad = \quad min(1, \frac{\hat{p}(z^*)}{\hat{p}(z^{(\tau)})}) \tag{2.15}$$

If the sample is accepted, than $z^{(\tau+1)} = z^*$, otherwise, $z^{(\tau+1)=z^{(\tau)}}$. It is worth noting, that if accepting sample z^* is going to increase probability $p(z)$, then the sample will certainly be kept. This procedure forms a basis of the *Metropolis* algorithm (Metropolis et al., 1953).

The Metropolis-Hastings algorithm (Hastings, 1970) is an extension of the basic Metropolis procedure to the cases where proposed distribution is non-symmetric. In this case, for a particular step τ, the sample z^* drawn from probability $q_k(z|z^{(\tau)})$ is

accepted with probability:

$$A_k(z*, z^{(\tau)}) = min(1, \frac{\hat{p}(z^* q_k(z^{(\tau)}|z^*))}{\hat{p}(z^{(\tau)}) q_k(z^*|z^{(\tau)})}) \tag{2.16}$$

where k denotes a set of considered transitions.

Further details of the procedures, along with proofs and derivations are not the main concern of this thesis and can be found in Metropolis et al. (1953); Hastings (1970). Detailed discussion of sampling procedures can be found in Bishop (2006).

2.3.2.2 Gibbs sampling

Gibbs sampling (Geman and Geman, 1984) is a special case of Metropolis-Hastings algorithm. This procedure involves replacing a value of one of the variables z_i from the current state $z = (z_1, z_2, ..., z_M)$ by a value obtained from a distribution conditioned on values of the remaining variables i.e. $p(z_i|z_{-i})$, where z_i is the i-th component of z, and z_{-i} are all remaining components.

This procedure is then repeated, with variables replaced in some particular order, or by choosing one of them to be updated at each iteration.

Although much simpler and faster than general Metropolis-Hastings algorithm, it is not applicable to all problems. Certain conditions have to be fulfilled in order for Gibbs sampling to produce samples from the desired distribution. Firstly distribution $p(z)$ has to be invariant of each sampling step individually, and as an effect of the whole Markov chain. This is, because while sampling from $p(z_i|z_{-i})$, the marginal distribution $p(z_{-i})$ is invariant due to value of z_{-i} being unchanged.

Secondly all conditional distributions have to be ergodic to ensure, that every point in space z can be reached in finite number of steps.

If these conditions are fulfilled the acceptance probability of Metropolis-Hastings algorithm is given by:

$$A(z^*, z) = min(1, \frac{p(z^* q_k(z|z^*))}{p(z) q_k(z^*|z)}) \tag{2.17}$$

$$= min(1, \frac{p(z_k^*|z_{-k}^*) p(z_k|z_{-k}^*)}{p(z_k|z_{-k}) p(z_{-k}) p(z_k^*|z_{-k})}) = min(1,1) = 1 \tag{2.18}$$

which effectively means that all the steps are always accepted.

Chapter 3

Data sets

3.1 Introduction

In this chapter we describe the datasets used throughout this thesis for the development and testing of models. The choice of our datasets has been limited by several factors. Firstly, the focus of this work on a multi-modal processing of eye movements dictates need for quality eye-tracking data. Secondly, confirming our claims requires the study of human behaviour over a variety of different tasks, rather than constricting the analysis to one phenomenon. Moreover the tasks need to involve a high level, contextual component that influences visual attention. In addition, for effective, incremental, and unbiased modelling it is necessary to have access to multiple datasets of varying complexity of both, contextual and visual components. Thirdly we focus on naturalistic scenes and try to avoid artificial stimuli such as visual arrays of symbols or letters. When possible we prefer photographic or photo-realistic scenes to clip-art composite images.

Finally we constrain our study only to relatively large datasets containing data collected over a variety of trials for several different subjects in order to extract amount of scan-paths allowing an effective modelling. We also preferr publicly available datasets rather than collecting our own data. This reduce the time cost and, more importantly, it allow easy comparison with other studies.

3.2 Visual Search datasets

In this study two different datasets based on *Visual Search* paradigm are investigated. The first dataset is from an experiment in which observers had to decide if the target

Figure 3.1: Example of photographs used in Torralba et al. (2006) study. The target object in all cases is *person*. Top row - scenes with target object i.e. person present, bottom row with target object absent.

was present or absent. The second experiment involved counting the number of occurrences of the target. These different paradigms results in different characteristics of the collected data.

3.2.1 Torralba's Visual Search set

The first dataset comes from Ehinger et al. (2009). It involves 912 urban photographs presented to 14 participants who were asked to decide if the target object was present or absent in each scene. The target object was a *person* in all trials, and is present in half of the images. Figure 3.1 presents examples of images from the dataset.

The eye tracking data was collected using an eye-tracker at 240Hz sampling rate. Images were presented at a resolution of 1024x768 pixels covering approximately 30 $\times$ 24 degrees of visual field. A Total of 12768 trials were collected, with an average length of 1.00 second $\pm$ 0.6 STD. The total of of 42545 fixations were collected, giving 3.33 fixations per trial on average.

3.2.2 Visual Count set

The second dataset comes from the work of Coco (2011). It involves a counting the number of occurrences of target objects in photo-realistic scenes (see figure 3.2 for example). A total of 72 scenes was presented to 24 participants, resulting in a total of 1738 captured trials. In contrast to the dataset of Torralba et al. (2006), the target object is always present. The number of targets - one, two or three - is a controlled factor.

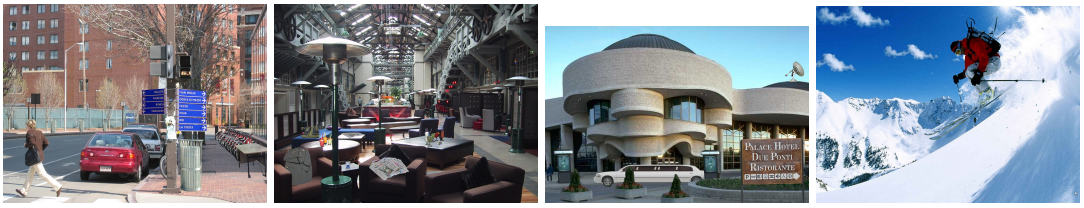


Figure 3.2: Example of scenes used in Coco 2011 studies. Targets on three images on the left are *man*, while for the last image on the right it is *goggle*.

Dataset	scenes	subjects	captured trials	fixations
Torralba et al. (2006)	912	14	12768	42545
Coco (2011)	72	24	1728	31420

Table 3.1: Comparison of visual search paradigm datasets used in this study. A clear difference in the characteristics of the data is visible - the average number of fixations is entirely different in both datasets being 3.33 for Torralba et al. (2006) and 18.18 for Coco (2011).

As a result, the characteristics of the collected scan-paths are different from those of Ehinger et al. (2009): average trial length was 4.84 seconds $\pm$ 3.96 STD, with 18.18 fixation per trial on average, and total of 54029 fixations collected. This difference is related to the necessity for more exhaustive scanning of the scene in order to perform counting. Table 3.1 presents side-to-side summary of both datasets.

The eye-tracking data was captured with the EyeLink II, head mounted eye-tracker, at 500Hz sampling rate. Images were presented at the resolution of 1024x768 pixels on a 17" screen, occupying approximately 30 $\times$ 34 degrees of visual field.

3.3 Language comprehension datasets

3.3.1 Language comprehension in Visual World

The major dataset used through this study comes from series of Visual World experiments described in Coco (2011) and is further referred to as the *Visual Arrays Visual World Paradigm* (VAVWP) Dataset. The dataset consist of three similar eye tracking experiments.

In each case, the visual stimuli consist of visual arrays containing three separate objects and two compositions of more than one simple object. Examples of the stimuli

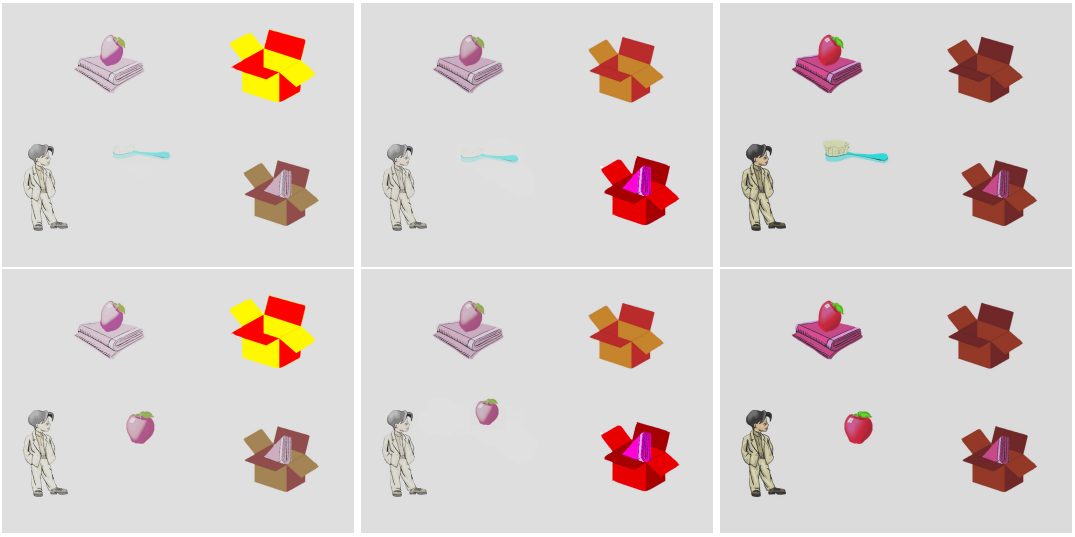


Figure 3.3: Example of visual arrays used as stimuli in VAVWP dataset. In each row: left - basic image, center and right - images with modified saliency

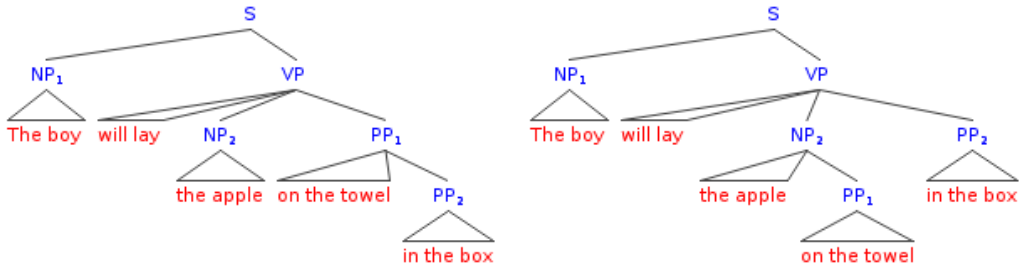
can be seen in figure 3.3. Each of the images comes in several versions. The basic version has a white background and objects with vivid colors. The other variations have the original background and one of the objects modified such that it is less or more salient (according to Walter & Koch's Saliency Toolbox) than the other objects.

The linguistic part consists of complex sentences played during the presentation of images. The sentences contain two prepositional phrases referring to objects in the image. Each sample comes in three versions - basic, with a continuous pronunciation of the words without any unnecessary pauses, and two variations with intentional breaks inserted either before or after the first prepositional phrase. Figure 3.4 presents an example of a sentence, its variations, and their possible syntactic interpretations.

The three mentioned experiments differ only in usage of the stimuli variations:

- Experiment 1 uses only images without modified saliency and sentences including variations;
- Experiment 2 uses images including those with modified saliency, and sentences without intentional breaks only;
- Experiment 3 uses all images and all sentences and their variations.

It is important to note that stimuli in Coco's materials are referentially ambiguous - that is the stimuli allow multiple interpretations, especially with respect to the grounding of words to objects. For example an object mentioned in the sentence might occur



The boy will lay the apple on the towel in the box.

The boy will lay the apple on the towel BREAK in the box.

The boy will lay the apple BREAK on the towel in the box.

Figure 3.4: Example of sentences used as stimuli in VAVWP dataset and their possible syntactic interpretations. The first sentence is ambiguous and both depicted parse trees are possible. In the case of the second and third sentences respectively the left and right trees are more likely.

	number of subjects	total trials	total fixations
Experiment 1	23	524	7985
Experiment 2	30	1046	25295
Experiment 3	32	1113	26763
Total	85	2683	60043

Table 3.2: Summary of the dataset from Coco (2011)

more than once - in case of sentences presented in figure 3.4, the *box* occurs two times in images from figure 3.3. Moreover it is these ambiguous objects, whose saliency is modified.

All the experiments follow the same setup - in the beginning a centre fixation cross is shown, followed by the presentation of an image for one second. After the preview the sentence is played, while the image continues to be displayed. After the sentence ends the image is kept on screen until a total of six seconds presentation time is reached. The eye tracking data is collected using EyeLink II eye tracker at 500Hz sampling rate. The images are displayed at 1024x768 pixels occupying 34 $\times$ 30 degrees of visual field. Table 3.2 presents a summary of the data collected in the mentioned experiments.

Even though the data is in some cases used as-is (i.e. sentences transcribed with

actual words, and scan-paths represented by series of coordinates and durations) it was necessary to compensate for the relatively large word and object vocabularies appearing in the stimuli. Therefore the sentences and collected scan-paths are encoded using the scheme described below.

The scan-paths are transcribed as a stream of objects found under fixated coordinates along with fixation durations. However, the objects identity is not used directly in transcription, but rather the semantic role of the word is used, which is inferred from the matching sentence. The set of labels used consists of *agent* - being an object that is the subject of the sentence, *patient* - being the object mentioned in the noun phrase of the sentence, and three additional labels for objects mentioned in the prepositional phrases and their competitors.

The sentences are transcribed in similar ways at various levels of granularity. The most general - from now on referred to as *PHRASES* - encoding follows syntactic structure of the sentence and attaches labels to each of the important sentence parts:

- NP1 - noun phrase containing *subject* performing an action,
- VP - verb phrase containing *predicate* performed by the subject,
- NP2 - noun phrase containing *patient* on which the action is performed
- PP1, PP2 - the first and second prepositional phrases

An example of sentences encoded with this scheme is presented below:

- PREVIEW [The boy]$_{np1}$ [will lay]$_{vp}$ [the apple]$_{np2}$ BREAK [on the towel]$_{pp1}$ [in the box]$_{pp2}$ POSTVIEW
- PREVIEW [The boy]$_{np1}$ [will lay]$_{vp}$ [the apple]$_{np2}$ [on the towel]$_{pp1}$ BREAK [in the box]$_{pp2}$ POSTVIEW

Notice the fact that the location of the intentional break does not affect the labels attached to the sentence parts, thus the disambiguation requires analysis of the whole sequence.

The second - *SEMANTIC ROLE* - encoding is very similar, however it uses position of intentional breaks to disambiguate the structure and attaches the following semantic roles to syntactic constructions:

- *agent* - being the subject of the sentence,

- *predicate* - being the verb denoting an action performed by the subject,
- *patient* - being the object on which the action is performed
- *locator* - being the prepositional phrase disambiguating another object - patient or target - by specifying its location.
- *target* - being the prepositional phrase denoting the final location of the action (such as *move, put*) given by predicate.

This encoding carries extra information allowing it to disambiguate the function of prepositional phrases:

- PREVIEW [The boy]$_{\text{agent}}$ [will lay]$_{\text{predicate}}$ [the apple]$_{\text{patient}}$ BREAK [on the towel]$_{\text{target}}$ [in the box]$_{\text{target locator}}$ POSTVIEW
- PREVIEW [The boy]$_{\text{agent}}$ [will lay]$_{\text{predicate}}$ [the apple]$_{\text{patient}}$ [on the towel]$_{\text{patient locator}}$ BREAK [in the box]$_{\text{target}}$ POSTVIEW

The final encoding does not use any syntactic information, and consist of transcription of sentences using part of speech (POS) tags:

- PREVIEW [The]$_{\text{det}}$ [boy]$_{\text{noun}}$ [will]$_{\text{modal}}$ [lay]$_{\text{verb}}$ [the]$_{\text{det}}$ [apple]$_{\text{noun}}$ BREAK [on]$_{\text{prep}}$ [the]$_{\text{det}}$ [towel]$_{\text{noun}}$ [in]$_{\text{prep}}$ [the]$_{\text{det}}$ [box]$_{\text{noun}}$ POSTVIEW
- PREVIEW [The]$_{\text{det}}$ [boy]$_{\text{noun}}$ [will]$_{\text{modal}}$ [lay]$_{\text{verb}}$ [the]$_{\text{det}}$ [apple]$_{\text{noun}}$ [on]$_{\text{prep}}$ [the]$_{\text{det}}$ [towel]$_{\text{noun}}$ BREAK [in]$_{\text{prep}}$ [the]$_{\text{det}}$ [box]$_{\text{noun}}$ POSTVIEW

Three additional tags are used in all three cases - *PREVIEW* denoting the initial period of silence before sentence is played, *POSTVIEW* being the time after sentence ends, and *BREAK* being the intentional break inserted between phrases.

The scan-paths are encoded as a stream of fixated objects. The objects are represented by labels derived from semantic functions of nouns referring to them in the sentence. Additionally, objects not mentioned in the sentence are encoded as *distractors*. These semantic functions are assigned based on the most likely syntactic interpretation of a sentence, rather than incremental partial parses.

The stream of labels is constructed by identifying the objects within which fixations fall. If a fixation does not fall on any object, it is assumed to be directed to a special object called *background*.

Figure 3.5 presents an example of encoding involving both image and sentence.

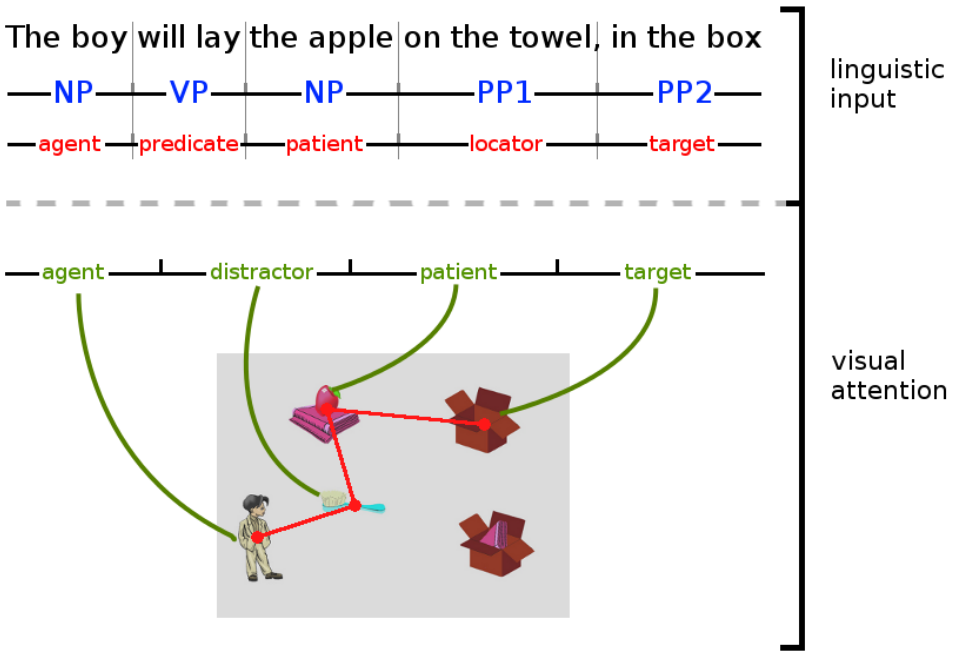


Figure 3.5: An example of data captured during experimental trials. The sentence is represented as a sequence of symbols denoting current phrase, or semantic function of the current word. The visual scan path is on the other side represented as sequence of fixated objects.

The woman sailed the yacht. The woman viewed the yacht.

The man putted the ball. The man detested the ball.

The woman parked the car. The woman cleaned the car.

Figure 3.6: Example of images and sentences used as stimuli in RVWP dataset.

3.3.2 Language comprehension in a Naturalistic Visual World

The second comprehension dataset - referred to further as Real Visual World Paradigm (RVWP) Dataset - used thorough this study - is very similar to the one described above. The difference lies in the stimuli used - naturalistic scenes instead of visual arrays. Unfortunately the linguistic part of the stimuli is much simpler and consist of shorter sentences involving only two objects (instead of five). An example of images and

sentences used can be seen in figure 3.6.

Encodings identical to those described previously are used. However, having a small number of objects mentioned in the sentences and a large number in the images, causes problems with transcription of scan-paths - the corresponding semantic role of the object is often not easy to define or the object might not be related to the sentence at all, while still competing for the visual attention. In our studies we used only trials where all objects mentioned in the sentences where present in the images.

3.4 Additional datasets

In addition to the datasets presented above, we have performed additional experiments and collected data required to study phenomena that are relevant for this thesis.

3.4.1 Object Naming

The first additional dataset contains data collected during an object naming experiment Clarke et al. (2013). The stimuli consists of 132 fully annotated images with a total of 2858 polygons with mean of 14.2 $\pm$5 STD and a median of 26 polygons per image. The annotations were produced by trained annotators with an accordance to pre-specified rules.

The images were presented to 24 subjects with a task explained by written instructions. Before each trial subjects were asked to fixate on a central cross. The image was than displayed for 5000ms, finished with a beep, after which the subjects were supposed to name objects presented in the scene. The set of objects that are allowed to be names was not restricted to annotated polygons. In fact the annotations are not accessible to the experimental subjects, hence each object, or its part present in the image could be mentioned.

The images were displayed on a 21" Multiscan monitor with resolution of 800 $\times$ 600 pixels. The eye tracking data was collected using Eyelink II eye tracker with 500Hz sampling rate. A total of 2904 usable trials were collected resulting in 88371 fixations.

3.4.2 Object interestingness judgement

The last dataset used in this thesis is entirely different from all datasets described above. The data was collected using Amazon Mechanical Turk. The set contains

Figure 3.7: Example of images used as a stimuli in Object Naming and Mechanical Turk interestingness judgement experiments. Typical responses to tasks of e.g. enumerating most interesting objects are: *cars, crossing, person* for the left, *bench, man* for the centre, and *barbecue, charcoal, chimney* for the right image.

100 images sourced from the Object Naming experiment described in section 3.4.1.

Each participant was presented with a batch of 30 images accompanied by a question asking them to mark and name one to three objects having specified features. Similarly to the naming experiment, any object present in the image can be selected. The marking is performed by clicking on the image. The location of the click is marked with a colour dot and text input pop-up is displayed to collect object name. The instructions ask for clicking in the centre of the selected object. Moreover the objects are to be marked at decreasing relevance to the subject of the question. Fifteen of these trials were associated with one of four critical questions, the rest were treated as fillers.

To avoid confusion only one of the critical questions is disclosed to a participant (i.e. the same question is presented in all critical trials). The critical questions are:

- *What is the most eye-catching object?*
- *What is the most salient object?*
- *What is the most interesting object?*
- *What is the most important object?*

The data of 175 participants was collected, resulting in a total of 2625 trials (i.e. image and question pairs). It is important to note that no eye-tracking data was collected during this experiment.

Figure 3.7 presents example of images from Object Naming and Interestingness Judgement datasets.

Chapter 4

Influence of contextual knowledge on human eye-movements

4.1 Introduction

Virtually every human activity occurs within a visual context and requires visual attention in order to be successfully accomplished (Land and Hayhoe, 2001). When processing a visual scene, humans have to localize objects, identify them, and establish the relations that hold between them. The eye-movements involved in these processes provide important information about the cognitive processes that unfold during scene comprehension (Henderson, 2003).

Studies of free viewing (e.g., Yarbus 1967; Einhauser et al. 2008) have shown that scan patterns on visual scenes can vary greatly between participants. On the other hand, the task that participants have to perform drives visual attention, resulting in fixated regions that are relatively consistent across participants both in search tasks (Torralba et al., 2006; Henderson et al., 2009) and in everyday activities (Pelz and Canosa, 2001; Hayhoe et al., 2003).

A number of models have been proposed to predict eye-movements during scene comprehension; they can be broadly divided into two categories. The first one consists of bottom-up models exploiting low-level visual features to predict areas likely to be fixated. Several studies have shown that certain features and their statistical unexpectedness attract human attention (e.g., Bruce and Tsotsos 2006). Moreover, low-level features are believed to contribute to the selection of fixated areas, especially for visual input that does not provide any useful high-level information (e.g., Peters et al. 2005). These experimental results are captured by models that detect salient areas of visual in-

put and predict attention in a bottom-up fashion. The best-known example is the model of Itti et al. (1998), which builds saliency maps based on color, orientation, and scale filters inspired by neurobiological studies of human vision. While there is evidence that saliency is predictive of eye-movement behavior (Itti, 2005), other authors have argued that this is merely a consequence of the fact that saliency is correlated with high-level properties that guide attention, such as objecthood (Castelhano and Henderson, 2007; Nuthmann and Henderson, 2010). Similarly Koostra et al. (2008) and Zhang et al. (2008) have shown that a range of others factors, including symmetry and Bayesian surprise, need to be taken into account when predicting fixation locations.

The second group of models assume that top-down supervision of attention contributes to the selection of fixation targets. Various types of top-down supervision have been observed experimentally. Humans show the ability to learn general statistics pertaining to the appearance, position, size, spatial arrangement of objects, and their semantic relationships. Chun and Jiang (1999) show that observers are able to temporarily learn contingencies between objects. Similarly, Green and Hummel (2006) show that perception is sensitive to the relative pose of pairs of objects. Hwang et al. (2011) demonstrate that observers tend to fixate objects that are semantically related in sequence.

A series of studies have also shown the importance of context in scene comprehension. Context not only provides information about scene layout scene and type (Schyns and Oliva, 1994; Renninger and Malik, 2004), but also about object presence, location, and appearance (see, e.g., Bar 2004; also Oliva and Torralba 2007, discuss the effects of context on object recognition in detail). Another important manifestation of context in scene understanding is contextual cueing: observers are able to associate the locations of target objects with arbitrary scene contexts, and use this information to speed up visual search when exposed to the same scene again (Brockmole and Henderson, 2006b,a; Brockmole et al., 2006). Furthermore, it has been shown that observers are able to extract low-level contextual information (scene gist) at very short exposures, without need for high-level visual processing (Castelhano and Henderson, 2007). This has inspired models that condition visual search on scene gist, such as the Contextual Guidance Model (Torralba et al., 2006), to which we will return below.

Whether visual memory is used during scene comprehension, as well as the exact form of such memory, is the subject of an ongoing debate in the literature. Several studies have indicated that visual search is memory-free (e.g., Horowitz and Wolfe 1998; Wolfe et al. 2000). Wolfe (1999) explains this result by proposing that vision

produces loose groupings of simple visual features such as the pre-attentive object files of Wolfe and Bennett (1997) or the proto-objects of Rensink (2000a), which dissolve upon the withdrawal of attention, meaning that visual search is memory-free.

But there is also a considerable amount of evidence for the opposite effect, i.e., the influence of visual memory in a range of search paradigms. For instance Gibson et al. (2000), Klein (1988), Klein and MacInnes (1999), and Takeda and Yagi (2000) all show that vision exploits information about which objects have been accessed within the same trial. Chun and Jiang (1998, 1999) show that memory can also be used across trials to guide attention. Also the contextual cueing effect discussed above is an example of visual search making use of information retained in memory across trials. In the context of the present paper, the study by McPeek et al. (1999) is particularly relevant. Using a visual search paradigm, the authors show that targets that match previously fixated targets are re-fixated more accurately and quickly than mismatching targets, indicating that attention is guided by short-term memory of visual features. Along the same lines, Maljkovic and Martini (2005) show that short-term memory can be used to explain effects of target frequency in visual search. In addition to this, memory effects have also been observed in other experimental paradigms (e.g., change blindness); for a more detailed discussion, refer to Hollingworth (2006), Shore and Klein (2000), or Woodman and Chun (2006).

The aim of this chapter is to explore the relationship between scene context and visual memory. Existing experimental and modeling studies dealing with context effects rely on an implicit form of memory, by assuming that participants remember, e.g., where objects are typically located in a scene (Torralba et al., 2006), or which object typically co-occur together (Hwang et al., 2011). We postulate a more direct link between visual memory and context. We test the hypothesis that the locations of the fixation that participant makes on a given scene can be predicted based on their fixations on directly preceding scenes. We present a model that stores fixation locations in memory, and compare its accuracy in predicting fixation locations to a model that relies on object context (Torralba et al., 2006). Our evaluation uses two data sets: an existing visual search data set from the literature (Ehinger et al., 2009), and a novel visual counting data set that we collected.

We also investigate the applicability of the general framework of saliency modulation to deal with incremental stimuli - such as data collected in Visual World Paradigm experiments.

4.2 Models of Context in Visual Attention

A number of models have been proposed to capture context effects on visual attention. A prominent example is Torralba et al.'s (2006) Contextual Guidance Model (CGM), which combines bottom-up saliency with a prior probability distribution encoding global scene information (gist). The central quantity computed by the CGM is the probability that a target object O is present at point X in the image:

$$p(O=1,X|L,G) = \frac{1}{p(L|G)} p(L|O=1,X,G)p(X|O=1,G)p(O=1|G) \quad (4.1)$$

Here, L is a set of local image features at X and G is a set of global features representing scene gist. The first term $\frac{1}{p(L|G)}$ is the saliency model. The second term $p(L|O=1,X,G)$ has the effect of enhancing the features of X that belong to the target object. The third term $p(X|O=1,G)$ is the contextual prior, which provides information about likely target locations. The fourth term $p(O=1|G)$ is the probability that O is present in the scene. The model is illustrated schematically in Figure 4.1. In Torralba et al.'s (2006) implementation of the CGM, the second and the forth terms are omitted, yielding:

$$S(X) = \frac{1}{p(L|G)} p(X|O=1,G) \quad (4.2)$$

This equation describes contextually modulated saliency $S(X)$ as the combination of bottom-up saliency and a prior on the likely location of the target, both conditioned on global features representing scene gist. These global features are computed by pooling local features over 4×4 non-overlapping windows; the resulting vectors are reduced using principal component analysis.

In following sections, we describe a model of visual attention that predicts fixation locations in visual search tasks. Our proposal is conceptually similar to the CGM, but the top-down modulation of saliency in our model is based on the memory of previously found targets, rather than on global scene properties. Moreover, we show that the knowledge of expected object locations can be learned incrementally, and that no prior is needed to achieve satisfactory results in predicting fixation positions. This avoids not only an expensive training phase, but also enables fast adaptation to different data sets, tasks, and experimental conditions. Additionally we show that combining both sources of knowledge (context and memory) enhances search performance.

The CGM has been extended to use additional sources of information i.e. target features detector (see Ehinger et al., 2009). However, we restrict our discussion to original model of Torralba et al. as the extensions are not directly related to the

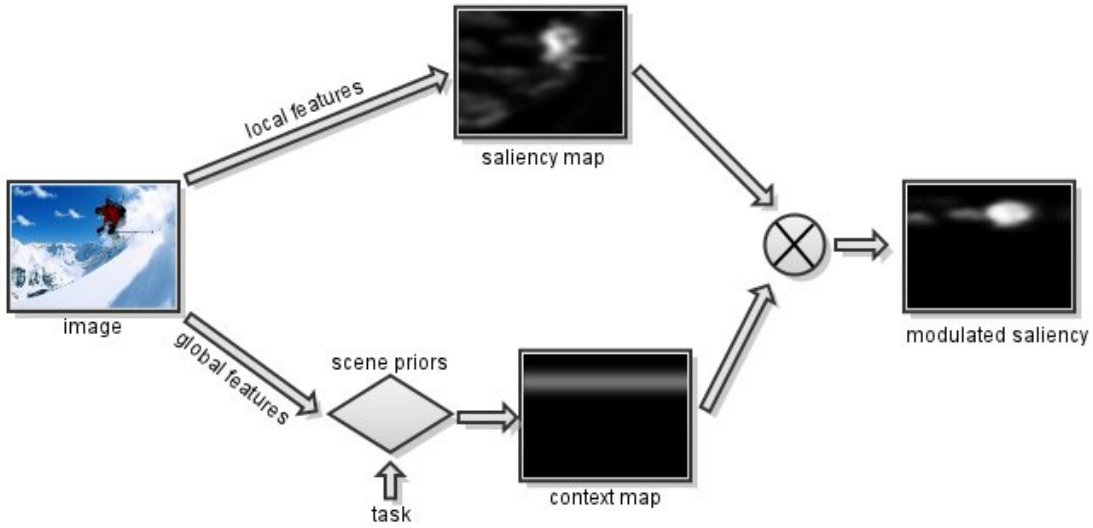


Figure 4.1: The architecture of the CGM. First, a saliency map is computed for the image. It is then modulated with a contextual prior conditioned on global scene features. The resulting map is thresholded to select the areas most likely to be fixated.

contextual guidance, and therefore cross-modal processing. Nonetheless we use the dataset of Ehinger et al. to perform more extensive evaluation.

4.3 Methods

4.3.1 Model Architecture

We propose the Memory Modulated Saliency (MMS) model of eye-movements in scene comprehension. Like the CGM, our model combines bottom-up saliency with a top-down estimate of likely target positions. In contrast to the CGM, our model does not assume any correspondence between global representations such as scene gist and human behavior. Instead, we assume that to estimate likely target positions, viewers rely on their memory of targets encountered in previous scenes. This information is then used to modulate a standard saliency map. The schematic architecture of the MMS model is shown in Figure 4.2.

Figure 4.3 presents an example of the computations performed by the model when fed a series of images. In the first step of each cycle, the saliency map of the image is calculated and modulated with the learned target position distribution. The resulting modulated map contains the model prediction for the fixation locations for this image.

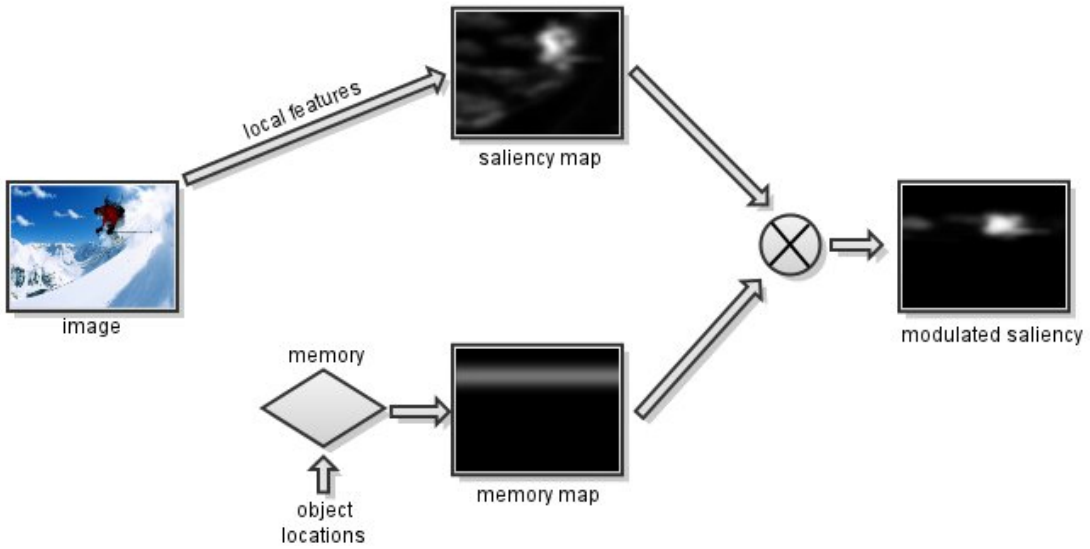


Figure 4.2: The architecture of the proposed MMS model. First, a saliency map is computed for the image. It is then modulated with a memory map estimated using fixations landing within the target objects or their center of mass on previously seen images. The resulting map is thresholded to select the areas most likely to be fixated.

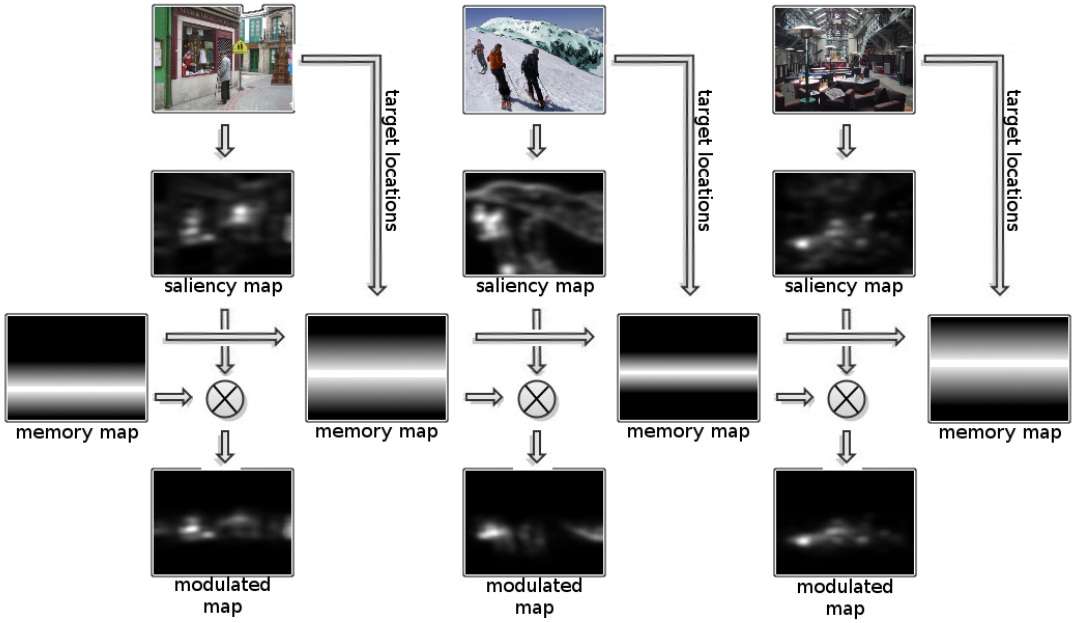


Figure 4.3: The computations performed by the MMS model. The incoming image is converted into a saliency map. The map is then modulated with a memory map computed based on target positions on previous images. resulting map is thresholded to select likely fixation locations.

In the next cycle, the distribution of target object locations is updated based on the fixations the participant made on the targets in the previous image. The resulting updated memory map is then used to modulate the saliency map for the current image, resulting in fixation predictions for this image. The actual fixations are then again used to update the memory map in the next cycle, and so forth.

4.3.1.1 Salience Map

We approximate saliency as the probability of the local images feature L in a given location based on the global distribution of these features (similar to Torralba et al. 2006):

$$p(L) \propto e^{-\frac{1}{2}[(L-\mu)^T \Sigma^{-1}(L-\mu)]} \tag{4.3}$$

Here μ is the mean vector and Σ the covariance matrix of the Gaussian distribution of local features estimated over the currently processed image. The local features are a set of Gabor filter responses computed over three color channels for six orientations and four scales, totalling 72 values at each position.

4.3.1.2 Memory Map

The top-down component of our model is implemented using memorized information, without access to image statistics or global scene representations. The MMS model learns a distribution over target object positions, and uses this distribution to modulate saliency. We make the simplifying assumption that this distribution is Gaussian.

A memory map is therefore represented as a single Gaussian distribution[1] whose parameters are estimated from the positions of the objects the participant fixated when viewing the n scenes preceding the current scene. Here, n is the memory depth, i.e., the number of scenes for which target locations are stored (e.g., a depth of three means that the three last scenes are considered). An additional simplification is that only the distribution of vertical positions is considered, while horizontal position assumed to be uniform. This is similar to an assumption made by Torralba et al. (2006).

For some images, no memory map can be estimated, because there are not enough target objects present in past scenes within allowed memory depth. This usually happens when the first few images in the experimental sequence are processed. A uniform distribution of target positions is assumed in this case.

[1] The histograms of target positions (see Figure 4.5) suggest that the distribution of target locations is slightly bimodal, so a modest improvement may result in employing a mixture of Gaussians instead of a single Gaussian.

4.3.1.3 Object Positions

The position of a fixated object can be stored in memory in a number of ways. A naive choice would be to use the center of mass of the fixated object as its position. This however does not capture the fact that objects are often relatively large, non-homogeneous entities, with fixations not always landing on the center of mass, or several unrelated fixations falling within an object's area. Moreover, this approach would not use the information provided by saccades and fixations directly. Hence the position of an object is approximated using following rules:

1. If a fixation falls within the object area, then the object position is approximated by the fixation coordinates.

2. If more than one fixation falls into the object area, than only the first one is taken into account, the other ones are discarded. This rule is justified by the fact that refixations of the target object only occur in a fairly small portion of the data (9.38% of trials for the visual search data and 12.60% of the trials for the visual count data).

3. If no fixations fall within area of a target object, then the fixations within one degree of visual angle are considered (with rules 1 and 2 modified appropriately). This is justified by the fact that the object are often small.[2] Furthermore, it rarely happens that no target object is fixated in a given trial (6.89% of trials for the visual search data and 5.54% of the trials for the visual count data).

4. If no position can be calculated using rules 1–3 then the object is assumed not to have been noticed by the participant, and thus discarded.

Once the object position has been approximated in this way, it is used to update the memorized distribution over target objects, as detailed above. After each update, the saliency map is modulated with the memory map to obtain the overall attention map.

Updates happen once per image based on the fixations on the target object in that image. If an image does not contain a target (which is the case for half of the images in the visual search data set), then no update is performed. If an image contains multiple targets, then all of them are used for the update. This situation occurs in the visual count data set.

[2]For the visual search data, the mean size of an object is 0.93° visual angle horizontally and 1.92° visual angle vertically. For the visual count data, the mean size is 1.77° horizontally and 3.90° vertically.

Figure 4.3 shows an example of how the various maps evolve over time in our model.

4.3.1.4 Memory Depth

An important questions regarding the computation of the memory map is what memory depth to use, i.e., how many previous fixations should be taken into account when estimating the distribution over target positions. In the experiments reported below, we manipulated memory depth by computing the memory map based on the three most recent fixations (MMS3), the ten most recent fixations (MMS10), or all previous fixations (MMSunrestricted). The memory depth of three was chosen as it provides a lower bound on what can be achieved by memorizing fixations locations: at least three fixation points are needed to estimate a Gaussian distribution over target locations. The memory depth of ten is based on the assumption that ten is the maximum number of fixations that can plausibly be held in human short term memory. MMSunrestricted is included to provide an upper bound on what a memory-based model can achieve. (Note that we will later also add MMSdual as a way of simulating category-specific memory.)

Note that assuming a Gaussian distribution over targets has potential limitation. People are able to capture and exploit more specific information such as the position of interesting areas or the spatial arrangement of objects (e.g., De Graef et al. 1990; Chun and Jiang 1998). Additionally, memory decay effects and the distinction between long and short term memory are not modelled by the MMS, even though they have been shown to have an effect on visual tasks (e.g., Davelaar et al. 2005). As mentioned in the Introduction, there is an ongoing discussion whether memory plays a role in visual search. However, it is important to note that previous studies have either been conducted on artificial stimuli (e.g., visual arrays), or focused on a particular phenomenon. Our aim, in contrast, is the more general one of investigating the role of memory as top-down supervision of low-level attentional mechanisms. For this, we believe, a simplified implementation of visual memory is sufficient.

4.3.1.5 Combined Model

We also investigate an extended version of the MMS model which combines the memory map with a contextual map representing prior knowledge as it used by the CGM. The modulation map M is constructed by a simple weighted mean of the memory map

MMS and a context map *CO* derived from the context oracle (see below for details on the context oracle). The value of the resulting map at position x, y is computed as:

$$M(x,y) = \omega \cdot CO(x,y) + (1-\omega) \cdot MMS(x,y) \tag{4.4}$$

Here, ω is a weight parameter determining the proportions at which the maps are combined. The architecture of this model is depicted in Figure 4.4.

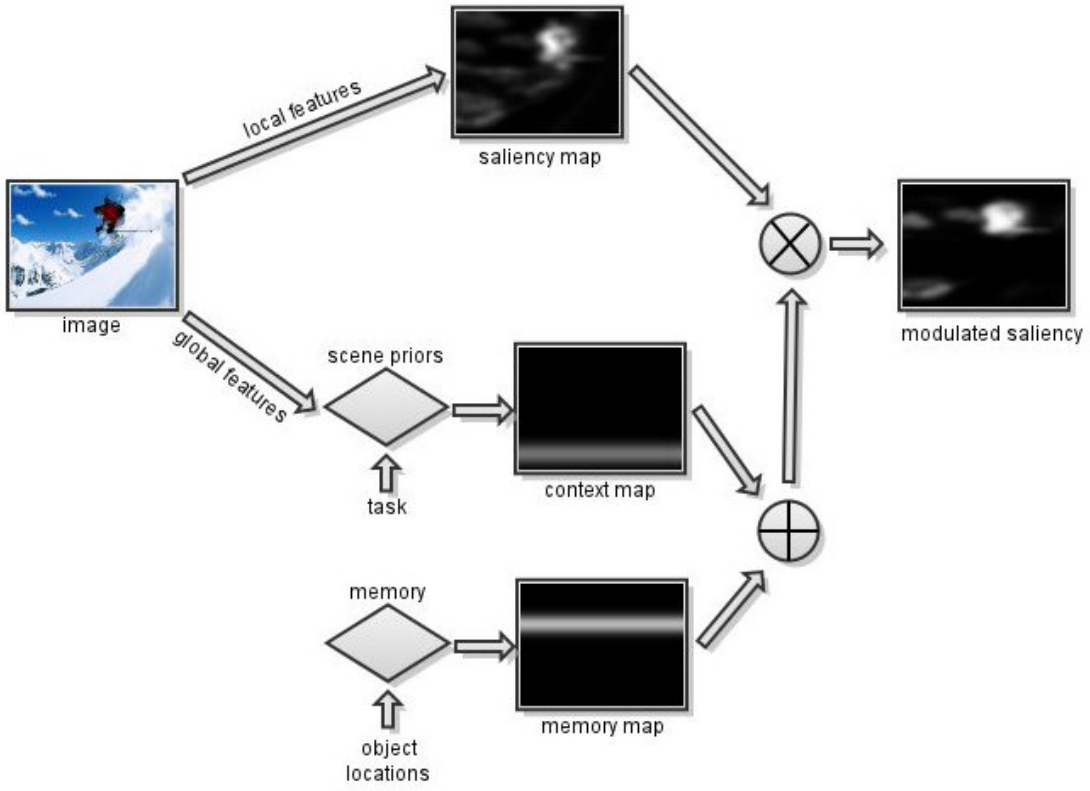


Figure 4.4: The architecture of the proposed joint model. First, a saliency map is computed for the image. It is then modulated with a map computed as weighted sum of the memory and context maps. The resulting map is thresholded to select the areas most likely to be fixated.

4.3.2 Visual Counting Experiment

4.3.2.1 Method

We evaluate the performance of the MMS model on eye-tracking data collected during a visual counting task. In this task, 24 participants were asked to count the number of occurrences of a cued target object, which was either animate (e.g., man, woman) or

inanimate (e.g., bin). The data set consisted of 72 photo-realistic scenes (both indoor and outdoor scenes), each containing one to three instances of the target object. The animate targets were all people, the inanimate targets were drawn from a wider range of categories; Figure 4.7 shows the frequency with which each target category occurred in the experiment.

A random order of the 72 scenes was generated for each participant (no blocking was used). Participants viewed each scene for as long as they liked, and then pressed one of three response buttons to indicate whether one, two, or three targets were present in the scene. Then the next scenes appeared; no feedback was provided.

The data was collected using a head-mounted eye-tracker with a sampling rate of 500 Hz. The images were displayed with a resolution of 1024×768 pixels, subtending a visual field of approximately 20 degrees.

4.3.2.2 Results and Discussion

The data set consists of 54,029 fixations collected over total of 1,738 trials. The average trial length was 4.84 seconds, with a standard deviation of 3.96.

A total of 23.58% of target objects were missed by the participants (i.e., not fixated in a given trial), this number is broken down by target category in Table 4.1. We find that more targets were missed when the cue was inanimate, and in scenes with a smaller number of targets. Regardless of the fairly large number of missed targets, the overall number of trials in which no target was fixated was low at 5.54%, therefore the estimation of the memory maps was possible at all times (except for the initial scenes where the map was assumed to be uniform, see above).

Table 4.1: Breakdown of missed targets by target animacy (rows) and number of targets in a scene (columns)

Cue	One target	Two targets	Three targets	Sum
Animate	4.08	0.63	0.50	5.21
Inanimate	7.19	7.07	4.11	18.37
Sum	11.27	7.70	4.61	23.58

4.3.3 Model Evaluation

4.3.3.1 Visual Search Data

In addition to the visual counting data described in the previous section, we also evaluated our model against the visual search data of Ehinger et al. (2009). In their experiment, 14 participants were asked to locate an animate target object, i.e., a pedestrian, in 912 naturalistic urban scenes, half of which contained the target. The data was collected using an eye-tracker with a sampling rate of 240 Hz, the images were displayed with a resolution of 800 $\times$ 600 pixels, subtending a visual field of about 24 $\times$ 18 degrees. This data set consists of 38,334 fixations.

4.3.3.1.1 CGM with Context Oracle We evaluated our model against a version of CGM that modulates saliency with a context map derived from a context oracle. The context oracle was introduced in connection with CGM to estimate an upper bound on what can be achieved with context-based models. The context oracle is based on manually annotated ground-truth maps, which were generated as follows. Participants are asked to mark on the y-axis the regions where the target object is likely to be found. Then these regions are then blurred using a Gaussian filter and aggregated over the different participants to obtain a single map for each image. We use the context oracle maps collected by Ehinger et al. (2009) for their data set, which are based on the context judgments of seven participants. For visual counting data, we generated our own context oracle maps, based on the judgments of five participants collected using the same procedure as used by Ehinger et al. (2009).

It is important to note that the CGM with context oracle can only serve as a approximate upper bound of CGM performance. It is not meant to estimate how much contextual guidance is possible in general. The context oracle is limited by the fact that each participant had to select a single y-axis location per target, even if there were multiple possible target locations. Effectively, the probability of not selected locations is estimated at zero; while this may be acceptable for some objects, it is unlikely to work for targets that can occur in a wide range of possible locations. Scene complexity is also potentially important: the assumption of a single location per scene is likely to work less well for complex scenes.

The parameters of the CGM model were set to the values reported by Torralba et al. (2006), i.e., the weight γ that trades of saliency and context maps was set to 0.05. In initial experiments we confirmed that this is an optimal value also for the data sets used

in the present paper. The same parameter settings were used for the MMS.

4.3.3.2 Performance Measures

In the Results and Discussion section below, we compare how the different models using **receiver operating characteristic** (ROC) curves. These curves plot the true positive rate of a model (also called hit rate) against its false positive rate. Our ROC curves are computed over all fixations a participant makes on a given image, as we are interested in how well the model predicts fixations in general, not just fixations on target objects. A true positive therefore is a fixation location correctly predicted by the model, a false positive is a fixation location incorrectly predicted by the model.

The models under investigation do not assume a fixed number of fixations per image; how many fixations a model predicts for an image depends on a threshold that determines what percentage of the image is selected for evaluation.[3] As the threshold is proportional to the false positive rate of the model, we will simply plot the threshold values on the x-axis of our ROC curves. In order to statistically compare model performance, we calculate the area under the ROC curve (AUC) of each participant. The AUC measures the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, and is equivalent to a Wilcoxon test of ranks, and closely related to the Mann-Whitney U-test (see e.g., Fawcett 2006). We submit AUC means to an ANOVA analysis, where we compare the performance of the different models pairwise, e.g., Saliency against MMSunrestricted.

In the visual counting data set, we also test the impact of target animacy on model performance. In line with the visual cognition literature (Fletcher-Watson et al., 2008), we expect our models to perform better on animate targets, as they are more quickly and accurately identified than inanimate targets, therefore exhibiting less variance in fixation behaviour. Note that the identification of inanimate objects is also complicated by the fact that they are more variable than animate objects, both in the terms of the range of object categories they belong to, and in terms of the positions at which they can appear in the image. We will return to this point in the section *Varying Memory Depth* below (see also Figure 4.8).

[3]Thresholding works by selecting the points with the highest model values until the threshold is reached. For example, a threshold of 10% on a saliency map means that we select the points with the highest saliency until we have selected 10% of the image. We then count how many of the fixations fall within these 10%. If we select 100% of image, we trivially predict all fixations correctly.

4.4 Results and Discussion

4.4.1 Distribution of Fixations

Figure 4.5 gives histograms of the vertical coordinates of the fixations in the two data sets. The histograms show percentages of all fixations (red lines) and percentages of fixations on the target objects (green bars). We find that these distributions are similar for both of the data sets. This finding is consistent with the hypothesis that visual attention is efficiently allocated to regions which are contextually relevant.

Alternatively, Figure 4.5 could also be explained by a central bias for both fixations and object locations, which has been reported in the literature (Tatler, 2007). This is a point to which we will return below, when we test a baseline model which remembers a random set of fixations, rather than storing the n most recent fixations. The random model matches the distribution of the fixations in the data set it is trained on; it therefore has an inherent central bias and should also pick up occulomotor biases that are present in human search behavior (Tatler and Vincent, 2009). Crucially, the random model does not use information about the order of fixations and therefore can serve as realistic baseline against which to compare the MMS model, which makes use of order information (see section Random Baseline below).

When we plot horizontal fixation positions for the visual counting data set (see Figure 4.6, left panel), we find a uniform distribution, which means that there is no general central bias for horizontal positions in this data set. For the visual search data, we find a bimodal distributions of horizontal positions, rather than a central bias (see Figure 4.6, right panel). The bimodality is an artifact of the experimental design which underlies this data set.[4]

[4]Ehinger et al. (2009) designed their stimuli as follows:

> For the target-present images, targets were spatially distributed across the image periphery (target locations ranged from 2.7° to 13° from the screen centre; median eccentricity was 8.6°), and were located in each quadrant of the screen with approximately equal frequency. (Ehinger et al. 2009, p. 950)

The fact that the authors placed target deliberately at the screen periphery explains the bimodality of horizontal positions in Figure 4.6 (right panel). There is only a weak bimodality in vertical positions in Figure 4.5 (right panel), which is probably due the fact that their target objects (which were always pedestrians) show a central bias vertically, which presumably counteracts the peripheral bias in the stimulus design.

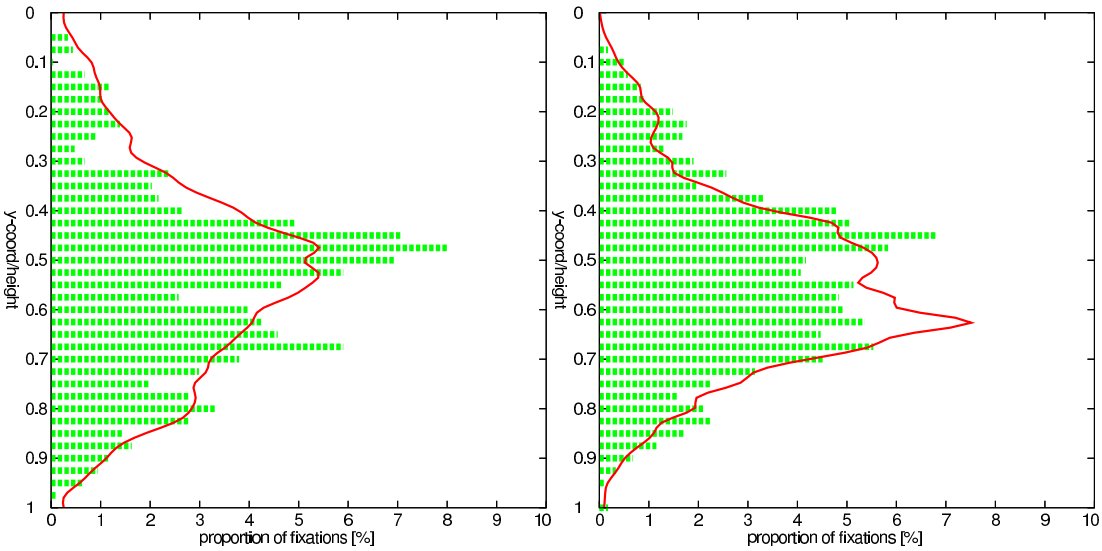


Figure 4.5: Histograms of vertical coordinates of fixations in visual counting (left) and visual search (right). The green bars depict percentages of fixations on the target objects; the red line shows percentages of all fixations.

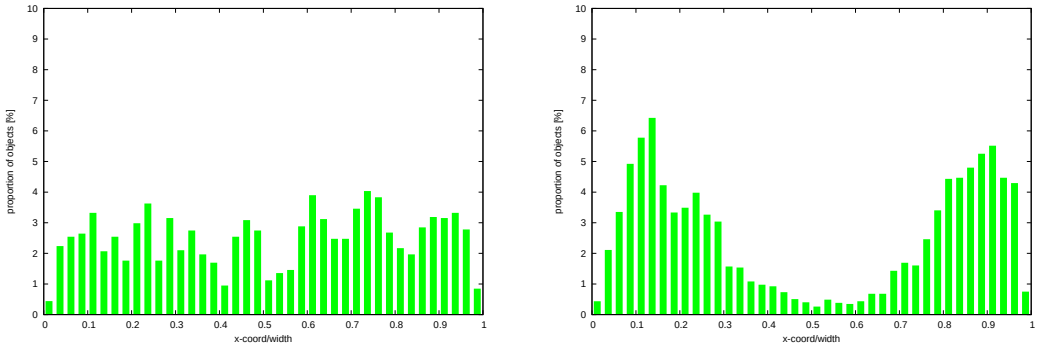


Figure 4.6: Histograms of horizontal coordinates of fixations in visual counting (left) and visual search (right).

4.4.2 Varying Memory Depth

Figures 4.9 and 4.10 show the ROC curves obtained by the different models for the two data sets. Overall, we find that the MMS models have a higher **hit rate**, i.e., proportion of fixations on target areas, than saliency in both data sets. This finding confirms that top-down knowledge is fundamental for model performance in goal-directed tasks, such as search. Crucially, we observe that even MMS models with small memory perform better than saliency.

In order to confirm this visual impression, we performed a pairwise ANOVA com-

paring the area under the RUC curve of the saliency-only model with the area under the curve of the MMS models (the AUC values are averaged over participants for both data sets, so the degrees of freedom are derived from the number of participants). The AUC values are summarized in Table 4.2.

For the visual search data set, we found a significantly larger area under curve for MMS3, the MMS model with a memory depth of three fixations, when compared to saliency ($F(1,13) = 27.8$, $p < 0.0001$). Also MMS10, with a memory depth of ten fixations, outperformed saliency ($F(1,13) = 192.8$, $p < 0.0001$). We obtained similar results for the visual counting data, where the area under the curve was not significantly different between saliency and the MMS3 model ($F(1,24) = 2.0$, $p > 0.1$), but it was larger for MMS10 compared to saliency ($F(1,24) = 26.6$, $p < 0.0001$).

Table 4.2: Performance of the models on the visual counting and visual search data sets. Given is the area under the ROC curve, averaged over participants (the table lists means and standard deviations).

Model	Visual search	Visual count
Saliency	75.33±1.10	80.91±1.68
MMS3	77.18±0.72	81.55±1.50
MMS10	79.60±0.89	83.22±1.47
MMSunrestricted	82.89±0.82	83.78±1.52
CGM	82.67±1.01	83.19±1.64
Random3	70.61±0.83	78.54±2.00
Random10	75.21±1.10	82.65±1.61

The difference observed between the two data sets is due to the larger variability in the visual counting task. The counting task used both animate and inanimate targets, while the search task only used one specific type of animate target (i.e., pedestrians). Furthermore, in the counting task, most animate objects belong to three frequent object categories, while inanimate objects belong to a larger number of categories, each of which only occurs once or twice (see Figure 4.7). It is also the case that animate objects are often located at the center and bottom part of the image, e.g., a pedestrian on a cross-walk, whereas inanimate objects can be found at a wide range of locations, see Figure 4.8. This source of variation is not present in the search data (compare Figure 4.8 to Figure 4.5). Moreover, the possibility of having multiple target causes

participants to inspect the scene longer than during the search task, which again increases the variability of visual responses.

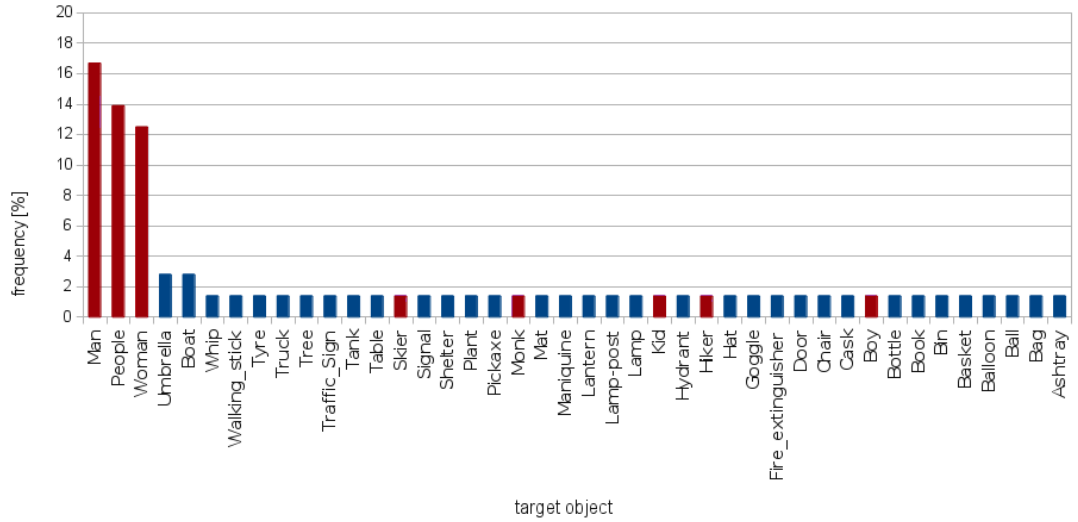


Figure 4.7: Frequency of different targets in the visual counting task. Marked red are animate objects while inanimate objects are blue. Note that most animate objects belong to just three categories, which while for inanimate objects are distributed over a larger number of infrequent categories.

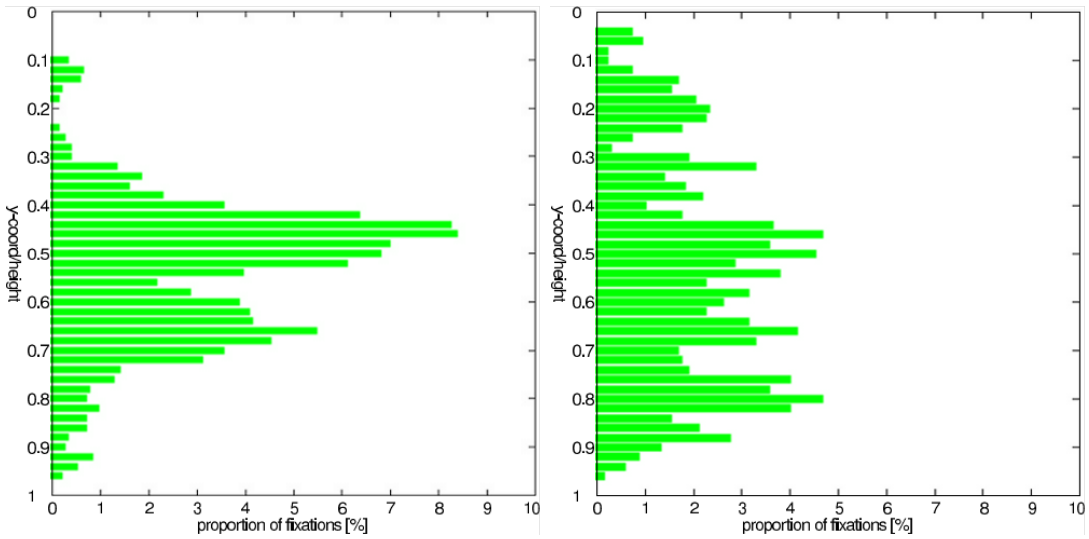


Figure 4.8: Distribution of vertical locations for animate (left) and inanimate (right) targets on the visual counting data. Animate targets are usually located at between half and two thirds of the image height, while inanimate objects are distributed more evenly across the image height.

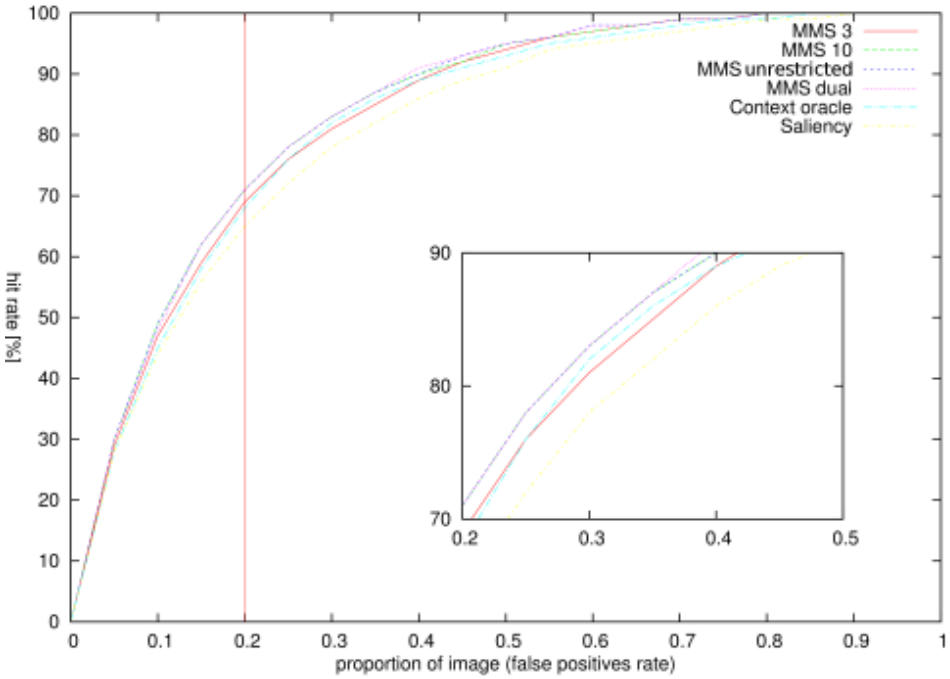


Figure 4.9: Prediction performance for the visual counting task for MMS with memory of three, ten, and an unrestricted number of fixations (MMS3, MMS10 and MMSunrestricted), MMS with a separate memory for animate and inanimate objects (MMSdual), the approximation of a CGM performance upper bound (CGM with context oracle), and the Saliency baseline. The curve is an ROC curve which plots true positives (hit rate) against false positives (proportion of image selected by the model). The red line marks the 20% threshold used by Torralba et al. (2006) in their evaluation.

When comparing the MMS models with the CGM with context oracle (i.e., the approximation of an upper bound of the performance of the CGM), we find that only MMSunrestricted, i.e., the memory model using all available fixations, is better than the CGM with context oracle, and only on the visual search data set ($F(1,13) = 5.4$, $p = 0.02$). We observe an improvement on the visual counting data set when we assume separate memories for animate and inanimate objects, i.e., MMSdual (to be discussed in more detail below). The performance of MMSdual is not statistically different from that of the CGM with context oracle ($F(1,24) = 2.9$, $p > 0.09$). Any model with a smaller memory performs worse than the CGM with context oracle on both data sets.

We repeated the evaluation using the position of the center of the mass of the target objects. In this analysis, the MMS did not memorize the fixation positions directly, but instead we computed the center of mass of the fixated object, and used this as the target

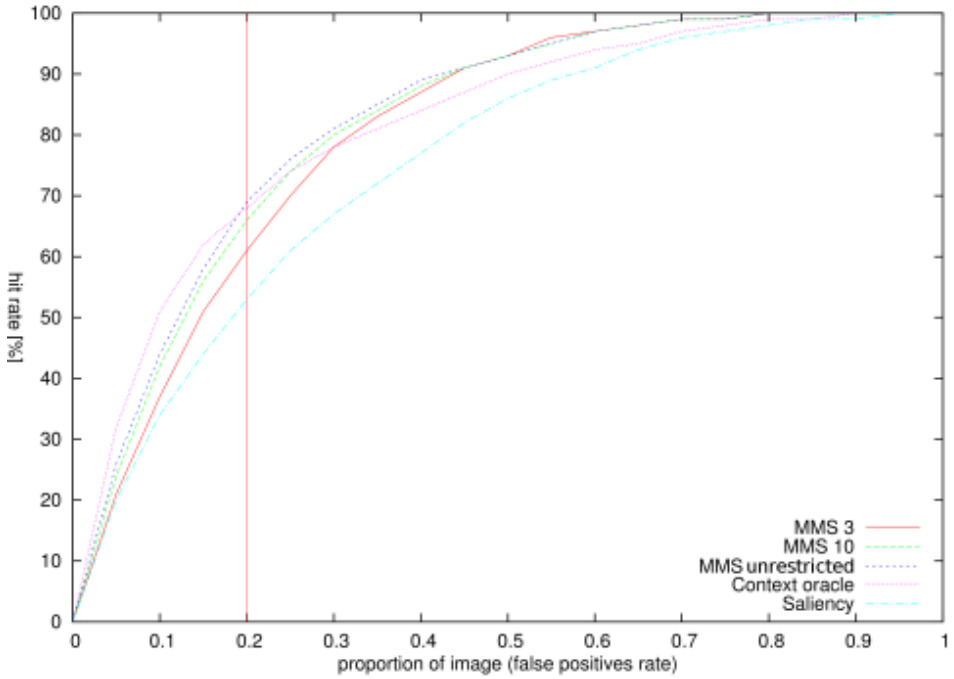


Figure 4.10: Prediction performance for the visual search task for MMS with memory of three, ten, and an unrestricted number of fixations (MMS3, MMS10 and MMSunrestricted), MMS with a separate memory for animate and inanimate objects (MMSdual), the approximation of a CGM performance upper bound (CGM with context oracle), and the Saliency baseline. The curve is an ROC curve which plots true positives (hit rate) against false positives (proportion of image selected by the model). The red line marks the 20% threshold used by Torralba et al. (2006) in their evaluation.

position for the MMS to memorize. This analysis was meant to simulate a situation in which no fixation data is available to the model, and it instead has to rely on object positions, just as the CGM does during training time.

This analysis revealed no difference in performance for visual count data. In the case of the visual search data, a difference was only observed for the smallest memory size, where using center of mass led to significantly improved performance (about 1.5% increase in AUC, $F(1,13) = 32.84$, $p < 0.0001$). Presumably, fixation data at a memory depth of three is fairly noisy, and this noise is smoothed out by using the center of mass of objects, rather than the fixation data directly.

More generally, the lack of a significant difference in most conditions between models using real fixation locations and the centers of mass means that the MMS does not have to rely on fixation data in order to update the memory. On more theoretical level, this result supports the finding of Nuthmann and Henderson (2010), who

show that the **preferred viewing location** of an object is close to its center of mass in naturalistic scenes; this is in turn predicted by the **cognitive relevance hypothesis** of Henderson et al. (2007) and Henderson et al. (2009).

4.4.3 Random Baseline

In order to provide a baseline against which to compare the MMS model, we also tested a version of the model that does not remember the n previous fixations, but a set of n randomly chosen fixations. This means that the baseline model does not have access information about the order in which the fixations occurred, but should capture general biases in fixation behavior and target locations, such as the central bias observed in the data (see Figure 4.5).

The baseline model was implemented by randomly scrambling the order of the fixations on the target objects. In doing so, we preserved the number of fixations per image, and fixations were randomized only within participants. The scrambled data set therefore has the same overall distribution of fixation locations as the original data set, and the same number of fixations is used to compute the memory map for each person/image combination.

We computed two versions of the random baseline model: Random3, which with a memory depth of three, and Random10, which a memory depth of ten (if we were to assume unrestricted memory then the random baseline would be identical to MMSunrestricted). The AUC values for these models for both the search and the counting data sets are displayed in Table 4.2. For the visual search data, we find that Random3 performs significantly worse than saliency alone, the worst predictor of fixation locations ($F(1,13) = 97.33,\ p < 0.001$), while Random10 is not significantly different from saliency ($F(1,13) = 0.1,\ p = 0.754$). On the visual count data, we again find that Random3 is significantly worse than saliency ($F(1,24) = 21.86,\ p < 0.001$), while Random10 is significantly better than saliency ($F(1,24) = 13.65,\ p < 0.001$) but significantly worse than the corresponding memory-based model MMS10 ($F(1,24) = 6.43$, $p = 0.0146$).

This set of results indicates that the performance of the MMS can not be attributed to general biases in fixation behavior and target locations, but is driven by the fact that the MMS uses the locations of the most recent target fixations to predict the location of the current fixation.

4.4.4 Dual Model and Combined Model

Model	Animate	Inanimate	All
Saliency	81.16±1.58	80.67±2.23	80.91±1.68
MMSdual	84.74±1.23	82.92±1.95	83.83±1.38
MMS10	84.61±1.51	81.84±1.90	83.22±1.47
MMSunrestricted	85.13±1.44	82.43±1.98	83.78±1.52

Table 4.3: The performance of the proposed models split by animacy of the target objects for the visual counting task. Given is the area under the ROC curve, averaged over participants (the table lists means and standard deviations).

Given the diversity of target objects in the Visual Counting dataset and different characteristics of object locations, it appears necessary to investigate effect of maintaining separate memory or contextual maps for each object type. However the dataset consists of trials with large number of different targets occurring infrequently (see figure 4.7). These targets can be divided into two categories. We will refer to them as *animate* (containing various instances of people), and *inanimate* (containing items and other non-living objects). A difference between animate and inanimate objects in terms of their typical location, justify evaluation of model performance on these categories targets separately. Table 4.3 provides the relevant AUC values. We observe that all models have a better performance on animate targets than on inanimate ones ($F(1,24) = 40.8$, $p < 0.0001$). This motivates the introduction of a dual memory version of the MMS model, which maintains separate memory maps for animate and inanimate objects, with a total memory span of ten fixations.

The difference between target objects can also be modelled by introduction of object features detector in a manner similar to extension discussed in Ehinger et al. (2009). However, this method does not takes into account differences in distribution of targets location. In fact feature detection does not model effects of contextual guidance, but rather bias of visual attention towards certain characteristics of the target object, and as such it is not a direct subject of this thesis.

This model (MMSdual) improves performance compared to a model with a single memory and a memory span of ten fixations ($F(1,24) = 3.9$, $p = 0.05$), but this is not sufficient to also outperform an MMS with unrestricted memory ($F(1,24) = 0.7$, $p = 0.39$). While this result is encouraging, it also raises questions regarding the level

of granularity that is appropriate for category specific memories. It is possible that the animate/inanimate distinction needs to be refined further, for example suing subdivisions such as human/animal for animate and artifact/natural object for inanimate. It seems plausible to assume that the MMS tracks a small number of object types and keeps separate memory maps for each of them. Furthermore, the granularity of the object types may be task-dependent. This is an issue that should be addressed in future research.

An analysis of the fixations generated by the MMS model and by the context oracle reveals that both models tend to predict fixations at different locations, in spite of their similar overall performance. Figure 4.11 presents the overlap o of predictions calculated as fraction of fixations found by both models for various threshold sizes:

$$o = \frac{|predicted(MMS) \cap predicted(CO)|}{|predicted(MMS) \cup predicted(CO)|} \tag{4.5}$$

where $predicted(\cdot)$ denotes set of fixations correctly predicted for a given model (MMS or context oracle), and $|\cdot|$ denotes set cardinality.

The relatively low overlap of the predictions for both models for smaller threshold values suggests that combining them should be beneficial. Indeed, we found that the simple joint model described earlier (see Figure 4.4 and equation (4.4)) section improves AUC values. The benefit of using a combined model is clear in the case of the visual search data, on which it achieves an AUC value of 86.01 (SD = 1.33) for a weight of $\omega = 0.6$. This AUC value is significantly better than that of the MMS model alone ($F(1,13) = 55.20$, $p < 0.0001$). In the case of the visual count data, the combined model achieves an AUC value of 84.26 (SD = 1.48) for $\omega = 0.4$, which however is not significantly different from the AUC value of the MMS model alone ($F(1,24) = 1.18$, $p = 0.28$). This can be explained by the fact that the overlap ratio for the two models is higher for the visual count data.

Overall, our results demonstrate that a simple model of visual search based on the memory of previous fixations can perform as well as, if not better, than a more complex model such as the CGM, which integrates bottom-up saliency with context information conditioned on global scene features.

It is also important to note that MMS model performance does not degrade on a visual count data set consisting of different scenes with a wide range of visual contexts. Instead, the MMS model still performs better than saliency and comparable to the context oracle on this data set. Moreover, we have shown that it is beneficial to combine both sources of knowledge: a model that includes prior contextual knowledge and

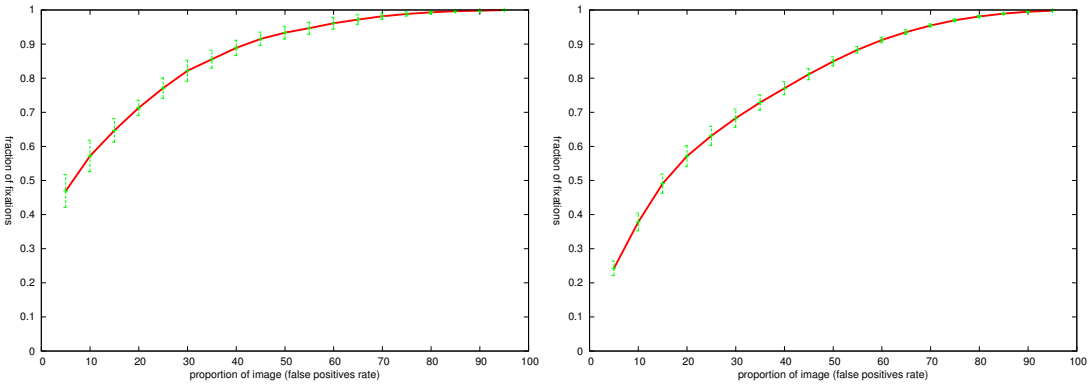


Figure 4.11: Overlap of fixations locations generated by MMS and CGM with context oracle calculated as the number of fixations found by both models over the total number of fixations predicted. Visual count data on the left, visual search data on the right. Note that this is not an ROC curve; rather, we plot the overlap of the models against the false positive rate.

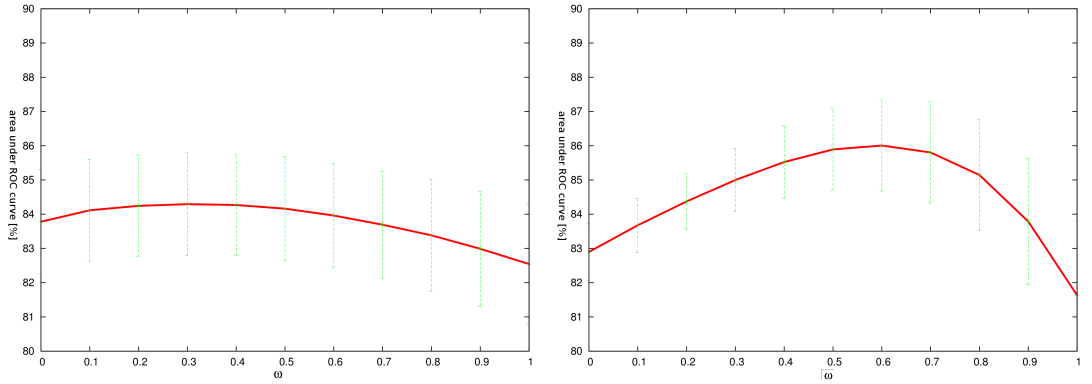


Figure 4.12: Performance of joint MMS/CGM model on visual count (left) and visual search (right) data. Note that this is not an ROC curve; rather, we plot the AUC achieved by the joint model against the relative weight ω of the contextual maps predicted by the two models used to modulate saliency.

memory of fixation locations showed improved performance, at least for the visual search data.

4.5 Evolution of context over time

Language comprehension - particularly language comprehension in Visual World - is a task with stimuli changing over time. Even though the image is static, the contextual (or rather linguistic) factor changes as the speech unfolds. This type of stimuli requires

construction of a model, that is capable of predicting interesting image regions at multiple consecutive time epochs, rather than - as with MMS and CGM - just considering initial snapshot of a scene.

The introduction of the temporal aspect into the saliency model, such that it is capable of producing maps of visually attractive regions that change over time is conceptually different from mechanisms such as *inhibition of return* (e.g. Itti et al., 1998), where the computed saliency map is constant, with fixated regions temporarily inhibited. In this concept we aim towards model, that allows *evolution* of a map combining both visual and non-visual features according to changes in the stimuli. In this particular case - stimuli consisting of an image and a sentence - the linguistic stream carries potentially important information, that shapes our understanding of a scene (e.g. by disambiguation of referents).

4.5.1 Model

We build a model around general framework established by CGM model - the saliency being modulated with a memory bound, however the bound itself changes over time following the structure of a sentence. Moreover the form of memory itself is different - instead of learning exact objects coordinates, the model learns probabilities of various types of the objects being fixated at given time.

The key part of the model is computation of a map M as modulation of saliency S with mask of objects T:

$$M(x,y) \quad = \quad S(x,y) \cdot T(x,y) \tag{4.6}$$

As in the MMS, we approximate saliency as the probability of the local images feature L_{xy} in a given location (x,y) based on the global distribution of these features:

$$S(x,y) \quad = \quad \mathcal{N}(L|\mu,\Sigma)^{-\omega} \propto e^{\omega\frac{1}{2}[(L-\mu)^T\Sigma^{-1}(L-\mu)]} \tag{4.7}$$

The purpose of $-\omega$ parameter is to transform the aforementioned probability - effectively a *fitness score* - into its inverse - *surprisal score*, and weight it with respect to temporal map T. During the evaluation value of ω was set experimentally to 0.5, although no significant differences were observed for values from range 0.01-2.0.

The mask T_t at time instance t and point $P = (x,y)$ is constructed as a sum of Gaussians centred at each object, with an amplitude proportional to proportion of fixations

at these objects during considered sentence segment estimated from a training set:

$$T_t(P) \propto \sum_{i=1}^{O} p_t(i) \cdot e^{-\frac{1}{2}[(P-l_i)^T \Sigma^{-1} (P-l_i)]} \tag{4.8}$$

where O is number of objects in the scene, $l_i = (x_i, y_i)$ is location of i-th object, and $p_t(i)$ is fraction of locations falling onto the i-th object during period t. Distribution p_t is estimated from the data as a fraction of time epochs in period t during which i-th object was fixated. Σ is controlling the spread of the Gaussian is chosen experimentally on a validation set.

It is important to note, that the mask can change over time. The extreme case when the sentence is assumed to be one period leads to modulation of saliency with overall probability of an object being fixated. In this model we equate period T to the course of phrase as the smallest recognizable unit of syntactic structure. Each of these phrases in a sentence is assigned its semantic role and fed into a prefix tree in order to form an unambiguous interpretation of a sentence at each time period. This allows us to reliably estimate $p_T(i)$ at each period, without necessity to parse the sentence during processing in order to obtain its interpretation.

4.5.2 Evaluation and discussion

The model was evaluated using 10-fold cross validation. In the experiment we used Visual Arrays and Real Visual World datasets with the semantic role based encoding (called SEMANTIC ROLE in chapter 3.3). This causes the periods at which the object map T is computed to be roughly equal to phrase lengths, with no additional variability related to syntactic ambiguity introduced. The proportion of fixations was calculated for each of the object types in each of the considered periods.

The significance of results was analysed with ANOVA. Averaged results are shown in table 4.4. The fraction of predicted fixation locations is calculated by thresholding the temporal map T at time period t corresponding to the onset of the fixation, such that areas with top 5% values is selected. In contrast to evaluation of CGM and MMS models all fixations collected at a particular trial are included in the evaluation - effectively hardening the task.

The reason to select only 5% of the image area is that visual arrays in general are very sparse, and the objects depicted do not occupy more than about 30% of space. Thus selecting larger regions leads to almost 100% hit rate due to all non-background areas being selected.

	saliency	object maps only	modulated saliency
VAVWP experiment 1	43.2%	34.9%	47.4%
VAVWP experiment 2	31.2%	23.3%	32.2%
VAVWP experiment 3	26.7%	29.4%	34.2%
RVWP dataset	22.5%	9.2%	23.5%

Table 4.4: Performance of the model modulating saliency with object reference maps updated on beginning of each corresponding phase. The results show overall fraction of fixations falling onto a target region. The target region is constructed by selecting 5% of image are with highest modulated saliency values at a given time point. The saliency and modulation maps were combined with equal weight.

The results clearly show that modulated saliency performs better than saliency on datasets with modulated intentional breaks (VAVWP experiments 1 and 3 with $F(1,22) = 21.45, p < 0.001$ and $F(1,31) = 24.82, p < 0.001$ respectively). Modulation of saliency reduces the effect of saliency modulation(VAVWP experiment 1), which disappears while the disambiguation is driven solely by saliency (VAVWP experiment 2, $F(1,29) = 2.29, p = 0.15$). These results are not a surprise, as they generally match the analysis of the human behaviour from Coco (2011).

The performance of the model and saliency is virtually the same ($F(1,7) = 0.53, p = 0.48$) for the naturalistic scenes dataset. One of the reasons is a construction of the dataset - only up to 2 objects are mentioned in the sentence (hence the reference maps are very weak and sparse), while large number of visually attractive objects not being referred to is present in the scenes.

It is interesting to see that using only object temporal maps - effectively selecting object centres with radius proportional to proportion of looks the object got in training data - performs well above chance level (of 5%), in contrast to static selection of objects centres (results not included in the table 4.4) which are at only 2-3% level.

Fine tuning of the model (e.g. by providing better method of constructing reference maps) might lead to better results, however the limitations of the models are obvious - the maps are recomputed for each phrase, while the behaviour of humans change continuously e.g. dynamics of scan paths change within each phrase after onset of nouns.

4.6 Summary

We presented a computational model that predicts fixation locations in visual search. Our approach is conceptually similar to the Contextual Guidance Model of Torralba et al. (2006), which combines saliency with scene gist and top-down context information about likely target positions. To obtain the context information, the CGM is trained on a large set of images with manually provided object labels. The Memory Modulated Saliency model that we propose, on the other hand, does not require offline training and does not involve the calculation of image or scene statistics. Instead, the MMS model keeps the last few fixations the participant made in memory, and uses them to predict likely positions of target objects.

The MMS model performs significantly better than saliency on two experimental data sets, demonstrating the benefit of memory for the prediction of fixation locations. An MMS model with unrestricted memory outperforms a context oracle (an upper bound on CGM performance) on visual search data, and achieves equal performance on visual count data. Unlike the CGM, the MMS does not require training data in the form of annotated images, but incrementally learns likely target positions. This means that the model can adapt easily to new data sets, tasks, and experimental conditions (unlike the CGM which is sensitive to the nature of the training data).

We also investigated whether a memory-based model needs to have access to fixation coordinates. We tested this by replacing the fixation coordinates with the coordinates of the center of mass of the fixated objects. We found that the performance of the resulting model is the same as that of the original version of the MMS that uses fixation coordinates. This means that fixations are not central to the model, they can be eliminated from it without degrading performance. All that the model requires is knowledge of fixated objects and their positions. It is therefore conceivable that the MMS can be trained in an offline fashion using images with object annotations, similar to the CGM, though the details of such a training scheme remain to be worked out.

We also demonstrated that a combined model that uses a weighted sum of the MMS memory map and the CGM context map outperforms the both individual models on the visual search data. This result indicates that a complete model of attentional guidance needs to combine features of both models. One important aspect of the CGM that is not present in the MMS is scene gist. In the CGM, the salience of a location in an image is conditioned on the scene gist (see equation (4.2)). It seems likely that integrating

gist would also be beneficial to the MMS: fixation locations (or alternatively, center of mass points) could be conditioned on gist in the same way. We leave this as an issue for future research.

Another potential limitation of the MMS model is that it requires a predictable, serial trial structure. It seems likely that memorizing fixation locations will only work if all the trials in a experiment are similar to each other (e.g., they are all search trials, or all counting trials, as in the two experimental data sets we tested). In an experiment in which different types of trials alternate, perhaps in an unpredictable fashion, having a memory of fixation locations in immediately preceding trials is likely to be less useful. However, it is conceivable that fixations memory could be conditioned on trial type, or that separate memories for different trial types could be maintained to solve this problem. (This would be similar to the dual memory model that stores animate and inanimate targets separately.)

We also found that the dual memory model, which stores the locations of animate and inanimate objects separately, outperforms a model with just one type of memory. If we assume that animate and inanimate objects differ in the their typical location in the scene, then storing their locations separately provides a restricted form of contextual guidance. It in conceivable to extend this approach and introduce separate memories for a larger number of categories. How many object categories are required is likely to be task specific, so a category-aware version of the MMS potentially should also include a way of learning which (and how many) categories need to be distinguished in a given task. Perhaps this learning could happen in an offline training phase just as in the CGM. This would then provide a less ad-hoc way of integrating the two models, a clear improvement over our combined model, which simply computes the weighted sum of the memory map and the context map.

An important conceptual difference between the two models is the type of learning that they capture. The CGM, by assuming an offline learning phase for target locations based on a large training set, effectively models how humans learn the typical positions of objects during childhood (or even beyond that for novel objects). The MMS, on the other hand, models short-term learning as it happens while human perform a specific task, and learn target locations based on where the targets where a few fixations ago. It seems that human behavior is driven by both types of learning; the two models should therefore be seen as complementary, pointing again towards the need for an model that integrates components from both the CGM and the MMS.

On a more theoretical level, our results provide an alternative explanation for the

tendency of experimental participants to only fixate contextually appropriate regions. In addition to using prior context information, participants seem to memorize likely target locations from previous trials, and use this information to guide their search on the current trial.

We have also discussed a model, constructed within the same framework, that is capable of dealing with incremental stimuli. The model outperforms saliency and a selection of fixation areas based on *Preferred Landing Position* (Nuthmann and Henderson, 2010).

The results show, that while applying contextual guidance models to Visual World Paradigm experimental data, it is beneficial to modulate saliency according to interpretation of the linguistic input at a considered time point.

The presented model dos not, however, fit into modern hypotheses such as Cognitive Relevance (Henderson et al., 2007) due to its pixel-based architecture. In addition the contextual bound is mere on effect of location biases, rather than arising of correspondence of objects in the visual scene and entities mentioned in linguistic input. As such no higher level knowledge, beliefs or associations can be easily modelled.

These drawbacks can be addressed with an object based models, assuming that attentional selection is driven by object's appearance and its contextual relevance to a performed visual task. This leads to an interesting issue of predicting which objects are more likely to be fixated during the course of a sentence, which is a problem discussed in chapter 5.

Chapter 5

Scanpaths in situated language comprehension

5.1 Introduction

Eye-tracking has become a widely recognized technique in psycholinguistics. The Visual World Paradigm (Cooper, 1974; Tanenhaus et al., 1995, VWP) allows for the study and understanding of various phenomena, giving insight into the temporal dynamics of attentional shifts (e.g. Altmann and Kamide, 1999, 2007; Knoeferle and Crocker, 2006, and many more).

The full understanding of the VWP data requires a framework explaining how language and vision interact to produce certain results. In this chapter we focus on eye-movements during situated language comprehension, aiming to provide a model capable of simulating natural scan-paths. We do not present a full cognitive theory of visual attention, instead we focus on building tools for modelling and understanding scan-paths, rather than - as it is commonly done providing an analysis of statistical properties such as proportion of looks. We hypothesize that scan-paths in linguistic tasks have structure going beyond simple correlation of fixations with objects mentioned in the linguistic stream. We believe this structure can be, to some extent, quantified, learnt from the data, and used to explain human behaviour.

Several models of VWP data have been proposed (e.g. Mayberry et al., 2009; Spivey and Dale, 2006; Kukona and Tabor, 2011), with some considering only the effects of a single word, and others aiming to integrate information coming from the entire part of a sentence seen so far (see e.g. Mayberry et al., 2009; Roy and Mukherjee, 2005). However, existing models focus on anticipation i.e. predicting the next

word or object of interest given the linguistic and visual context, rather than capturing eye-movements directly.

Moreover we believe several aspects of the analysis and modelling are given little attention, despite their importance. These include consideration of eye movement dynamics at finer than sentence or word level, modelling eye movements at global and local level, and consideration of individual differences between subjects.

We aim to develop a set of tools that will be able to explain such behaviour on an individual basis, rather than provide insight into collective trends only. Our ultimate goal is to show that behaviour on a fixation-to-fixation basis is, in certain tasks, is consistent enough across experimental subjects to perform studies more detailed than general statistical analysis.

5.1.1 Time in models of visual attention

The majority of the existing models of visual attention neglect the existence of various aspects of visual attention, such as: time, the order of fixations, or the incremental character of the stimuli.

For example, models of saliency, such as that of Itti et al. (e.g. 1998), are static in their nature, providing only a map of visual attractiveness over two dimensional space. Similar models based on saliency, such as Contextual Guidance Model (Torralba et al., 2006), or Memory Modulated Saliency presented in chapter 4, neglect the existence of time.

The original work of Itti et al. (1998) includes an attempt to model temporal aspects of attention, by means of *winner-take-all neural networks*, and *inhibition of return* mechanism to select a sequence of fixation points and simulate human scan paths. This solution does not however change the static character of the underlying saliency, and in fact, does not model incremental scene interpretation.

These types of models are therefore very hard to extend to accommodate synchronous stimuli with their temporal and incremental aspects. Our initial attempt presented in chapter 4.5 faces the same limitations. Even though this model performs incremental updates of the context map according to the current interpretation of the linguistic stimuli, the solution is ad-hoc and capable only of updating the saliency map, rather than generating scan-paths.

A completely different approach to modelling has been proposed in the model described in Mayberry et al. (2005). In this case the incremental processing of syn-

chronous input is used as a fundamental assumption. The model however does not attempt to model attention, or generate scan-paths, but rather focuses on anticipation - that is prediction of the next referent mentioned in the linguistic stream (i.e. word) based on the current interpretation of the sentence and visual context. It is however important to remember that prediction of words does not equate to prediction of attentional shifts, which might not only occur several times during each word, but also might involve objects that are competing for attention (e.g. during ambiguity resolution), or are not currently mentioned in the linguistic stream (e.g. re-fixations to previously accessed objects).

Similarly the *FUSE* model of Roy and Mukherjee (2005) is based on the assumption of incremental processing of synchronous stimuli. This model maintains probability distributions over objects being fixated as a processed sentence unravels. These probabilities are used to create speech recognition systems assisted with information from the visual context.

The closest work to our goal is that of Kukona and Tabor (2011). Their model predicts the shifts of attention that might be interpreted as scan-paths directly. The described implementation based on attractor artificial neural networks is however handwired i.e. the neurons' activation weights are set manually rather than learnt from experimental data. This makes it applicable only to small, simple datasets, where weights for all the network neurons can be easily calculated or assigned experimentally. More diverse and complex datasets would require a larger numbers of neurons (hence more parameters) and more complicated calculations to be performed, rendering it impossible to be done manually. As a result, the model does not generalize and scale well to new datasets. Moreover it does not consider attentional shifts at a level finer than that of a single word.

5.2 Modelling human attention in linguistic tasks

5.2.1 Problem Formulation

The prediction of scan-paths during speech comprehension is the main focus of this chapter. Even though it seems to be possible to extend models for location prediction with temporal components, they still lack the ability to predict scan-paths directly. Extensions such as the mentioned above *winner-take-all* networks of Itti et al. (1998) are in fact an ad-hoc addition and do not explicitly utilize any information about structure

of human scan-paths and their correspondence to the linguistic input.

Results presented in studies such as Coco (2011) show that scan-paths and language are coordinated beyond simple grounding of nouns to visual objects. Rather than that, more complex dynamics can be observed such as a competition of two objects for attention that changes its characteristics over time.

We will leave aside the problem of predicting exact coordinates and durations of fixations, focusing on sequences of fixated objects. We will also assume visual factors - such as saliency - to be less important for a moment. This leads to the key idea behind further modelling - treating generation of scan paths as an alignment of two sequences: one encoding the sentence and the other being the fixated objects.

Let each sentence be represented as a sequence of observed symbols $S = s_1, s_2, ..., s_T$, where s_T is a symbol (e.g. word) at time t.

The problem of finding a sequence of fixated objects $\hat{O}$ can be than expressed as:

$$\hat{O} = \arg\max_{O} P(O|S) \tag{5.1}$$

This probability is intractable directly as it would require computing all possible sequences O, but can be rewritten using Bayes' Rule:

$$P(O|S) = \frac{P(S|O)P(S)}{P(O)} \tag{5.2}$$

and for a given set of priors $P(S)$ the probability depends only on the likelihood $P(S|O)$.

In the particular case of scan-path prediction and generation investigated through this chapter, the sequence S stands for a sentence using labels representing parts of speech of individual roles, or derived from the semantic role of nouns within each phrase (for details see 3.3.1). For example, sentence *The boy will lay the apple on the towel BREAK in the box* can be represented in one of following ways:

- PREVIEW [The boy]$_{\text{np1}}$ [will lay]$_{\text{vp}}$ [the apple]$_{\text{np2}}$ [on the towel]$_{\text{pp1}}$ BREAK [in the box]$_{\text{pp2}}$ POSTVIEW
- PREVIEW [The boy]$_{\text{agent}}$ [will lay]$_{\text{predicate}}$ [the apple]$_{\text{patient}}$ [on the towel]$_{\text{patient locator}}$ BREAK [in the box]$_{\text{target}}$ POSTVIEW
- PREVIEW [The]$_{\text{det}}$ [boy]$_{\text{noun}}$ [will]$_{\text{modal}}$ [lay]$_{\text{verb}}$ [the]$_{\text{det}}$ [apple]$_{\text{noun}}$ [on]$_{\text{prep}}$ [the]$_{\text{det}}$ [towel]$_{\text{noun}}$ BREAK [in]$_{\text{prep}}$ [the]$_{\text{det}}$ [box]$_{\text{noun}}$ POSTVIEW

referred to as *PHRASE*, *SEMANTICROLE* and *POS* representations. It is possible to use more fine-grained representations - such as letters or phonemes - which form the

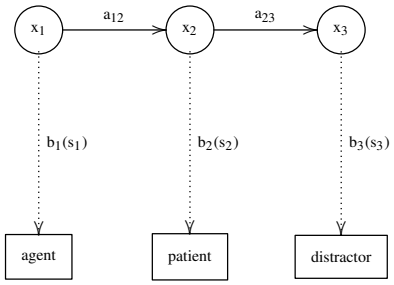


Figure 5.1: An example of Markov Model M depicting the process of generating sequence S of fixated objects.

basic units of a language - words. They are also not directly related to the meaning or syntactic structure of a sentence, besides being subunits of words, which are associated with scan-path characteristics (see e.g. Coco, 2011; Altmann and Kamide, 1999; Tanenhaus et al., 1995). We believe that it is not necessary to use such artificial, low level representations of a sentence in our work.

The sequence O on the other hand, represents scan-path as sequences of fixated objects. Each object is represented by a symbol derived from a phrase containing the noun referring to this object: subject, noun phrase (patient) and prepositional phrase 1 and 2. More detailed information, including further examples and their visualization, can be found in chapter 3.3.1.

5.2.2 Prediction of fixation locations based on POS/Semantic role

We will approach the problem by constructing series of Hidden Markov models (HMMs) - a widely used technique for recovering sequential relationships.

In HMMs a sequence of observed symbols (i.e. a sentence) is generated by a Markov model (see figure 5.1). A Markov model can be seen as a *Finite State Machine*, with transitions occurring always at a fixed time unit t. At each transition a state j is entered and an observed sentence symbol (e.g. word) is generated from probability density $b_j(s_t)$. The transitions from state i to j are also probabilistic and depend on probability a_{ij}.

The joint probability that S is generated by model M is calculated simply as a product of transitions probabilities and output probabilities:

$$P(S,X|M) = a_{12}b_2(s_1)a_{22}b_2(s_2)a_{23}b_3(s_3)... \tag{5.3}$$

where X is the underlying state sequence. In practice only observed sequence S is known, while state sequence X is hidden. The required joint probability can be ap-

proximated by considering only the most likely state sequence $X = x(1), x(2), ..., x(T)$:

$$P(S, \hat{X}|M) = \max_X \{a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(s_t) a_{x(t)x(t+1)}\} \quad (5.4)$$

Given a sufficient set of training examples the parameters of the model (transition and emission probabilities) can be computed efficiently.

In practical applications it is often the case that a collection of HMMs is built, and the sequence is generated by concatenating sequences generated by more than one of them, thereby maximizing likelihood. Such an approach is widely used in continuous speech recognition, where HMMs are trained to generate phonemes instead of whole words, and are in turn concatenated to produce these words. In the described model, the produced HMMs are counterparts to the symbols encoding the sentences.

5.2.2.1 Evaluation

For training purposes the time-course of the recorded trials is divided into a set of non-overlapping windows. Each of the data streams is represented by one symbol for each of the time-windows. These symbols represents the objects (in the case of fixations stream O) or the part of the sentence (in the case of linguistic stream S) that was present for the longest time during the considered period.

Two such encodings are possible - one with input and output sequences directly reflecting the human behaviour that is a synchronous sequence has length of the scan-path and the sentence symbols are repeated as appropriate. Although very simple, this encoding has one very important drawback - the decoding phase requires prior knowledge of the number of fixations for each part of each processed trial.

Assuming the simplest case with only one fixation for each sentence part [1] in the decoding phase the HMM is conceptually equivalent to the model of Mayberry et al. (2005), who use a simple recurrent network to align words and objects.

We evaluate our models in terms of average distance between real and generated scan-paths calculated using modified Scan-Match toolbox (see chapter 6.4.3 for details). The results obtained for this model during initial evaluation that are presented in table 5.1 as **windowed HMM (phrase)**, show that HMMs do not predict human-like scan-paths - not only are the Scan-Match distances between real and synthesized sequences much greater than agreement between subjects, but also the sequences themselves contain no realistic dynamics at all. For example, the models predict attentional

[1] By sentence part we would refer to a word or any continuous subsequence of words in the sentence forming a syntactically coherent entity such as phrase

Model	VAVWP exp 1	VAVWP exp 3
windowed HMM (50ms)	0.97	0.96
windowed HMM (phrase)	0.97	0.96
Agreement between subjects	0.36 ± 0.011	0.32 ± 0.010

Table 5.1: Average Scan-Match distance between scan-paths produced by discussed HMMs, and human behaviour on Visual Array Visual World Paradigm (VAVWP) datasets presented in chapter 3.3.1.

changes at word onsets, while these might also occur multiple times within each word leading to more complex patterns.

However, investigation of the sequences shown below reveals that semantic roles of some objects (as described in section 3.3.1) can be align with the corresponding phrases.

For example, the sentence:

$[\text{The boy}]_{\text{np1}}$ $[\text{will lay}]_{\text{vp}}$ $[\text{the apple}]_{\text{np2}}$ $[\text{on the towel}]_{\text{pp1}}$ BREAK $[\text{in the box}]_{\text{pp2}}$

is aligned with following sequence of objects:

$[\text{SUBJECT}]_{\text{SUB}}$ $[\text{SUBJECT}]_{\text{PRED}}$ $[\text{PATIENT}]_{\text{PAT}}$ $[\text{PP1 competitor}]_{\text{PP1}}$ $[\text{PP2}]_{\text{PP2}}$

This confirms the basic finding of VWP experiments, i.e., objects are fixated when or shortly after they are mentioned. However, when syntactic or visual ambiguity occurs (i.e., more than one object can correspond to a semantic role), the HMM is prone to errors caused by variability in the data, and predicts either the correct object or its direct competitor at a given time frame, depending on the number of fixations each receives in the training set (see the object predicted at the first prepositional phrase of the sentence).

The alternative approach is to divide the time-course of a trial into a set of equally long time epochs. The synchronous sequence than consists of current part-of-sentence and the object fixated for each time epoch. In our experiments the epoch lengths were experimentally set to 50ms. The results of such alignment, called **windowed HMM (50ms)** in table 5.1, are however equivalent to the results obtained with the HMM described earlier. This is due to the nature of the Maximum Likelihood (ML) estimation of the HMM transition and emission probabilities - simply put, the Viterbi algorithm recovers the most probable (in terms of ML) objects and transitions effectively outputting the most fixated object for each sentence part.

An additional observation can be made - these sequences do not have the dynamics of a real human behaviour such as a competition of objects for the attention, refixations etc.

Through the rest of this chapter we will not discuss the problem of predicting timing and duration of fixations. Instead we will focus solely on finding a natural sequence of fixated objects.

5.2.3 Generation of scanpaths using Markov-Chain Monte Carlo methods

The main problem associated with prediction of scan-paths using the HMMs mentioned in previous section is that the probability of a fixation is conditioned on one symbol (e.g. word) of a sentence only. It was suggested however, by various Visual World Paradigm studies, that the dynamics of eye-movements can change depending on the current interpretation of a sentence. It is therefore important to consider whole chunks of a sentence seen so far. Moreover HMMs do not capture higher order correlations - that is emitting a symbol is essentially independent of emitting any other symbol in the sequence.

To solve the problem described above, we base our predictions on a sequence of probability distributions that describe human behaviour during each sentence part. The MCMC is proposed in order to enable utilisation of complex probability distributions taking into account variety of dependencies between fixations, sentence, and other possible factors. Estimating parameters and maximising such probability functions is often computationally infeasible. A comparison of architectures of the simple HMM model and the proposed alternative is depicted in figure 5.2.

The idea of predicting probability distributions rather than fixated objects themselves is not new and was applied in the *FUSE* model of Roy and Mukherjee (2005). The *FUSE* model does not however predict scanpaths, but uses probabilities as priors for language models used to improve speech recognition.

The work of Kukona and Tabor (2011) also uses the idea of utilizing probability distributions (in a setting similar to the object masks as described in section 4.5) to predict attention shifts. However, the attentional shifts can occur only when the probabilities change - effectively once per sentence part - which is conceptually equivalent to the HMMs described earlier and the model of Mayberry et al. (2005).

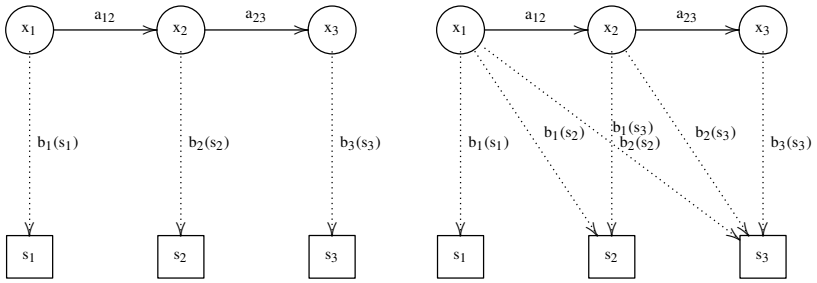


Figure 5.2: Comparison of graphical representation of models discussed in sections 5.2.2 (left) and section 5.2.3 (right). The first model generates the sequence of symbols representing fixated objects $O = (o_1, ..., o_N)$ based only on the current symbol in the sentence, while the other model considers the whole part of a sentence seen so far.

5.2.3.1 Sampling scan-paths using a bi-gram model

In the model described here, we align a set of probability distributions with an interpretation of a sentence after seeing each sentence symbol - e.g. partial parses - using an extension of an HMM. Other solutions are possible (including those used in studies mentioned above such as SRN, or even maintaining fixed probability lattices for each phrase type). This is not discussed further as an exact solution used is not crucial for understanding the proposed approach.

Having the sentence interpretations aligned with the probability distributions, the model generates the sequence of fixated objects with the Markov-Chain Monte Carlo (MCMC) approach. The simplified generative process for the scanpath is shown below.

- scanpath = empty
- for each sentence part as s_p
  - sample length l from discrete probability $P(l|I(s_1...s_p))$
  - for i = 1 .. l
    - sample object o_i from $P(o_i|I(s_1...s_p), o_{i-1})$
  - append subsequence $o_1...o_l$ to scanpath

The initial scanpath is then re-sampled using the Metropolis-Hastings method, with the probability of the whole sequence defined as:

$$
\begin{aligned}
P(O|S) &= P(o_1 ... o_N | s_1 ... s_M) && (5.5)\\
&\propto P(len(O) = N) \cdot \\
&\quad \prod_{i=1}^{M} P(len(O_{I(s_1 ... s_i)}) = n_i) \cdot \\
&\quad \prod_{i=1}^{M} \prod_{o_j \in O_{I(s_1 ... s_i)}} P(o_j) \cdot P(o_j | o_{j-1}, I(s_1 ... s_i)) && (5.6)
\end{aligned}
$$

Where $O = (o_1, ..., o_N)$ is a sequence of fixated objects (scanpath), $S = (s_1, ..., s_M)$ an encoded sentence, o_i and s_i i-th object and sentence part respectively, O_{s_i} is a subsequence of O at sentence part s_i, while N and M are the total length of scanpath and sentence. Finally, $I(s_1 ... s_i)$ is the interpretation of a sentence after seeing the i-th symbol. It represents a partial parse of this sentence - essentially introducing incremental language processing into the model. Our implementation of the model is based on constructing an infinite HMM implemented using a prefix tree. Each state in such an HMM corresponds to a partial parse tree. This approach is less powerful than a real, dedicated incremental parser. Using such a parser, or other methods, is obviously possible, but not necessary to handle datasets used in our experiments. The exact method of parsing used is not an essential part of the model, therefore we only assume that it is capable of recovering distinct states in syntactic processing of a sentence.

This formulation explicitly considers a syntactic interpretation of the whole part of a sentence seen up to the moment the fixation is generated (addressing the first problem discussed earlier) and puts constrains on the length of the scanpath and its subparts (addressing the limitations of the models capable of generating transitions only at discrete time points). It is also important to note that this formulation essentially enforces a bi-gram language model of the scanpath symbols, as the probability of sequence element depends solely on its predecessor. The importance of this assumption is discussed further, in later sections.

5.2.3.2 Scan path re-sampling with shallow-parsing

The dependency of the model presented above on a bi-gram relationships between fixated objects, is a main limitation to its extensibility. Estimating more powerful n-gram models would require not only considerably larger training sets, but also appropriate smoothing, and dealing with n-grams spanning over different sentence parts.

As in the previous model, the initial scan-path is sampled using the same generative procedure. The subsequent re-sampling is however modified - using the Metropolis-Hastings procedure, the acceptance of the sample is not based directly on bi-gram probability. Instead a generated sequence is analysed at a higher level, by grouping fixations into sub-sequences. We hypothesise that such coherent, and repeatable groupings occur within real scan-paths and their presence is driven by linguistic processing.

We propose to extract the fixations groupings using syntactic chunkers as used in Natural Language Processing. Using the syntactic chunkers, the probability of the generated sequence for the MCMC acceptance test is approximated by the probability of the parse calculated as:

$$P(O|S) \approx P(C|I,O) \tag{5.7}$$

Where O is a sequence of fixated objects (scanpath), $I = (I(s_1), I(s_1,s_2)...I(s_1...s_n))$ is a sequence of sentence interpretations after seeing each new symbol, X is the sequence of HMM states corresponding to the sequence of tags C assigned to fixated symbols, $a_{x_i x_{i+1}}$ transition probability between state x_i and x_{i+1}, and $b_{x_i}(o_t)$ emission probability of fixation o_i in state x_i.

The syntactic chunking is defined as a tagging problem, where fixations are assigned a label indicating whether they belong or not to a larger group of consecutive fixations called *chunk*. Each group consists of beginning, end and internal fixations, with each fixations belonging to at most one such group.

We perform tagging with HMM-based model. Each state of the HMM corresponds to a possible label, and the parse is defined as a sequence of transitions corresponding to the sequence of fixations that maximizes the probability:

$$P(C|I,O) \approx \max_{X=(x_0,...,x_n)} a_{x_0 x_1} \prod_{t=1}^{T} b_{x_t}(o_t) a_{x_t x_{t+1}} \tag{5.8}$$

where $C = c(0),...,c(T)$ is the sequence of labels (effectively chunk tags), corresponding to a state sequence $X = x(0),...,x(T)$, $a_{x(t)x(t+1)}$ and $b_{x(t)}$ transition and emission probabilities, while F and I is the scan-path being parsed and sequence of sentence interpretations.

This approach is meant to favour sampling of longer structures of fixations which occur within the scan-paths, even if probabilities of the bi-grams they consist of are not improving. At the same time no large training is required, and the chunkers are robust to unseen sequences without necessity for smoothing. They can also handle fixations

that appear to be anomalous - such as off-object fixations - simply by not assigning them to any chunk.

Further details of syntactic chunking and the model described above referred to as *Basic HMM chunker* can be found in chapter 6.

5.2.3.3 Experimental evaluation

The evaluation was done using materials from Coco (2011) described in section 3.3.1 including all three encodings of the sentences. Experiment 2 of this dataset was excluded from evaluation, as it was used for development of the models and, what is more important, contains no sentences with intentional breaks, hence less linguistic variability.

It was expected that PHRASE and SEMANTIC ROLE encodings should yield the same results given sufficiently large HMM aligning the sentence with probability distributions used by the sampler in the second stage, as such HMMs would be capable of recovering and representing distinct interpretations of sentences regardless of their less informative encoding. To ensure this, the HMM was constructed by building prefix trees representing sequences of partial parses of a sentence - in practice this approximates an HMM with an infinite number of states.

The POS encoding is intuitively more flexible, with the potential to produce more fine grained alignments, at a cost of reducing the amount of data available to estimate probability distributions for each HMM state.

The models are evaluated in terms of an average distance between real and generated scanpath. The distance is computed using *Needleman-Wunch* algorithm (see e.g. Cristino et al., 2010, for applications to scan path comparison). The results obtained with 10 fold cross-validation over 100 runs are presented in table 5.2. The results are tested for statistical significance by performing pairwise ANOVA on average similarity scores within each fold obtained with compared models.

As expected the HMMs used for alignment are able to deal with encoding using chunks and produce results not significantly different to those while using pre-parsed phrases (ANOVA $F(1,54) = 0.32, p = 0.57$). However, the results obtained using parts of speech are significantly worse and slightly counter-intuitive. Deeper investigation of the probability distributions reveals that the data for some of the sentence parts (e.g. determiners) is very sparse. This is because the probabilities are estimated by counting number of fixations onsets in each segment. For example, a new fixation occurs during only about 2% of the determiners of the subject's noun. Such sparsity is a significant

Model	Needleman-Wunsch distance	
	VAVWP exp 1	VAVWP exp 3
HMM alignment (Phrases)	0.97 ± 0.001	0.96 ± 0.001
Sampling (Interpretation POS)	0.43 ± 0.021	0.32 ± 0.015
Sampling (Interpretation Sem. Roles)	0.36 ± 0.011	0.27 ± 0.011
Sampling (Interpretation Phrases)	0.36 ± 0.011	0.27 ± 0.011
Sampling (Symbol Phrases)	0.39 ± 0.013	0.37 ± 0.010
Sampling + Chunking (Interpretation Sem. Roles)	0.35 ± 0.012	0.35 ± 0.006
Agreement between subjects	0.36 ± 0.011	0.32 ± 0.010

Table 5.2: Results for the prediction of sequences of fixated objects - lower distance is better. HMM alignment denotes a baseline model that predicts most probable object for each part of sentence; Sampling denotes an extended model that aligns sentence with probability distributions for different encoding schemes (POS, PHRASE, or SEMANTIC ROLE) and alignment methods (current symbol only or current interpretation). Chunking denotes models which use shallow parsing for calculation of scan-path sequence probabilities. The results obtained with bi-gram based models using phrase and chunk representations of sentence are not significantly different from each other and agreement between subjects on VAVWP 1 dataset. HMM alignment and sampling based on POS representation are however significantly worse, while the introduction of chunking significantly improves the results. The trend is similar to that of the VAVWP 3 dataset, with the exception of sampling being significantly better than agreement between humans.

problem resulting in severe misestimation of HMM parameters.

It is interesting to see that aligning fixations o with sentence symbols S_i based on probability $P(o|S_i)$, rather than with partial parses based on $P(o|I(S_1...S_i))$ leads to results worse than agreement between humans even though it addresses the limitations of the models discussed earlier and allow multiple fixations to occur during each sentence part. This shows that current interpretation of a sentence (hence grounding of words to visual objects) affects the human behaviour. It is an important finding, as it shows that interpretation of the linguistic stimuli augments the visual processing, which might potentially lead to different incremental interpretations of the scene.

Another interesting observation is that the discussed models perform better on experiment 3 of the evaluation dataset than on experiment 1. This can be explained by consistent visual cueing towards certain objects on stimuli in experiment 3, while the stimuli in experiment 1 consists of equally salient objects presenting more ambiguity to be resolved, and as a result, has higher variability of human responses.

It is visible from the results that generating sequences that are within the range of differences between humans is possible with the presented model based on bi-grams. However, the similarity measure used considers only the amount of differences rather than their type. For example, substituting a fixated object with one of remaining ones yields the same distance even though some of them are more likely to appear in this position than the others. Indeed investigation of the distribution of sequences of three objects fixated in a row (as in tri-gram model, see table 5.3) shows that some structures present in the human data are virtually not existent in simulated data, which is dominated by other sequences emerging from projection of bi-gram probabilities to longer sequences. The differences in distributions can be explained by the fact that

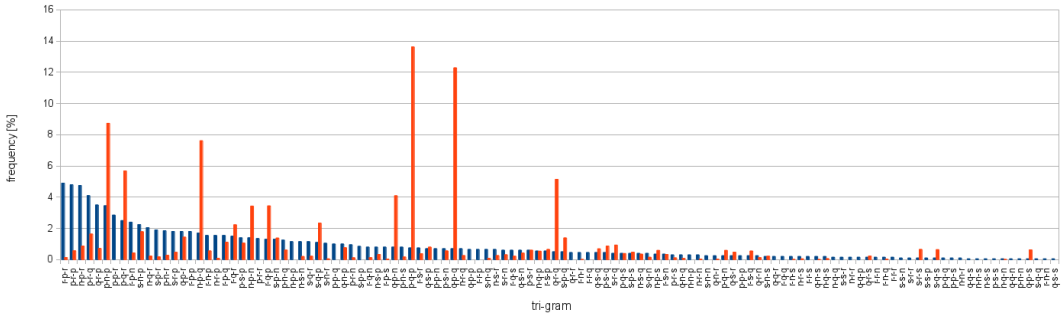


Figure 5.3: Distribution of tri-grams in human (blue), and simulated (orange) scanpaths for experiment 1 of VAVWP dataset. The chart presents the trigrams sorted according to frequency in human data. Letters denote objects referred in following parts of sentence: s - subject, n - noun phrase (patient), p - prepositional phrase 1, r - prepositional phrase 2, q - competitor of object referred in prepositional phrase 2, as discussed in chapter 3.3.1

the described bi-gram based method does not have any control of behaviour at spans longer than two objects. Secondly the sampling favours bi-grams with greater probabilities which causes different behaviour at larger time spans, effectively ruling out some sequences that occur less frequently - as a result, only 80 out of possible 125 tri-grams occur in simulated data, while 118 can be found in recorded eye-tracking tri-

	Human data	Simulated scan-paths
r-p-r	4.90%	0.14%
p-r-p	4.80%	0.57%
n-p-r	4.75%	0.87%
p-r-q	4.10%	1.64%
q-r-p	3.50%	0.72%
p-n-p	3.45%	8.73%
p-p-r	2.85%	0.00%
p-q-r	2.50%	5.68%
r-p-n	2.40%	0.43%
s-n-p	2.25%	1.79%
p-q-p	0.75%	13.62%
q-p-q	0.70%	12.28%
n-p-q	1.70%	7.61%

Table 5.3: Example of trigrams with their frequency in human and simulated scan-paths. Letters denote objects referred in following parts of sentence: s - subject, n - noun phrase (patient), p - prepositional phrase 1, r - prepositional phrase 2, q - competitor of object referred in prepositional phrase 2, as discussed in chapter 3.3.1

als. Moreover in simulated data, the histograms of bi-gram frequencies closely follows Zipf distribution, while in real data it is entirely different (see figure 5.3).

This problem is not surprising as bi-gram models are not capable of handling tri-gram dependencies in the sequences. The expressiveness of bi-grams results in the inability of the model to simulate human behaviour at spans longer than two consecutive fixations, even though the global proportion of looks towards different objects are matching experimental data. It can be solved by building more expressive n-gram models, however it is infeasible due to sparseness of available data (see chapter 6.2 for details).

The model presented in section 5.2.3.2 overcomes this problem by integrating syntactic chunks as described in section 6.2. Using the chunkers enables the model to

recover and utilise sequences longer than two fixations. This leads to generation based on longer, more coherent progressions of fixations that might be seen as equivalent to whole phrases in language processing.

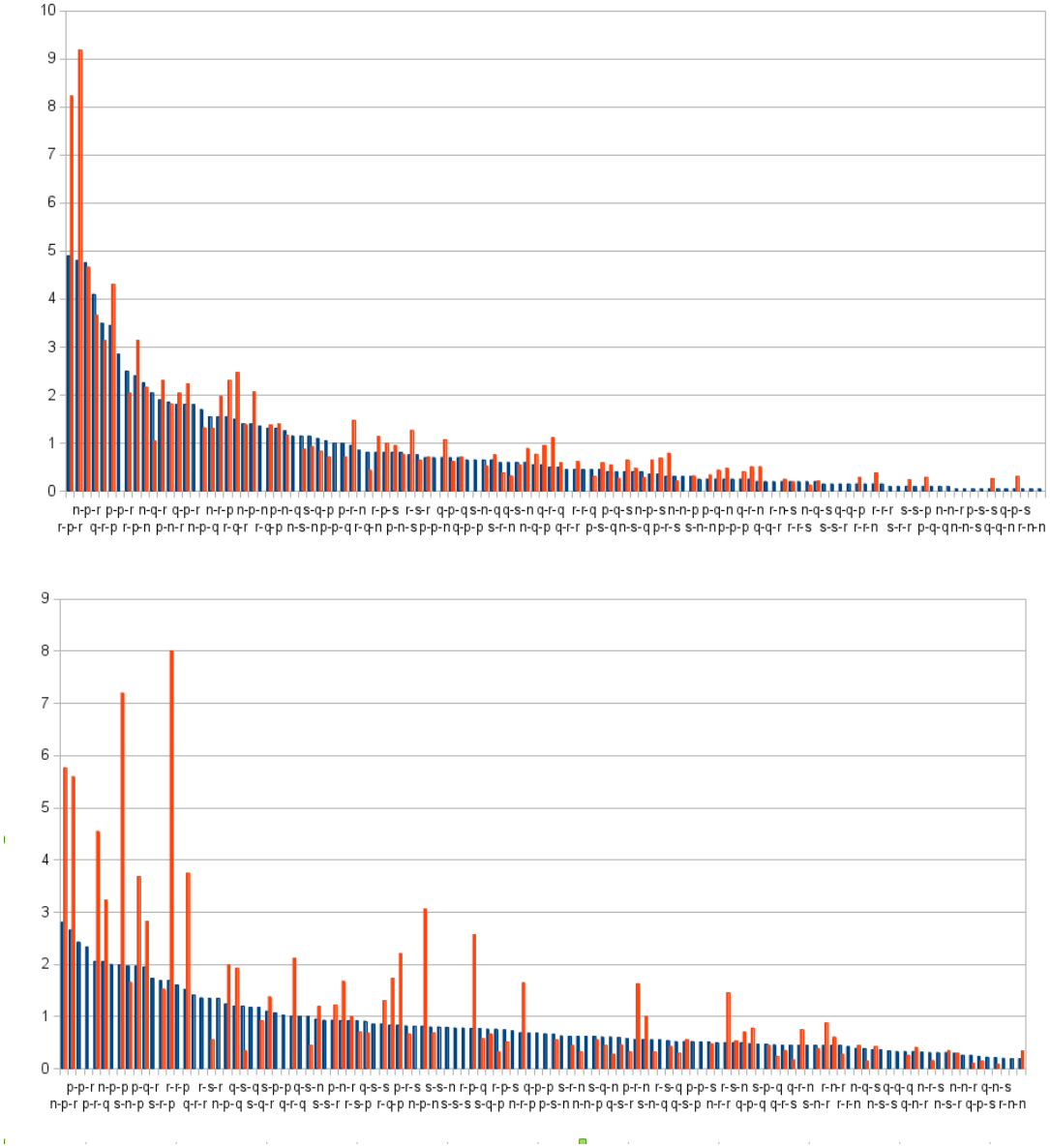


Figure 5.4: Distribution of tri-grams in human (blue), and simulated (orange) scanpaths. Top chart corresponds to experiment 1, while bottom to experiment 3 of VAVWP dataset. Letters denote objects referred in following parts of sentence: s - subject, n - noun phrase (patient), p - prepositional phrase 1, r - prepositional phrase 2, q - competitor of object referred in prepositional phrase 2, as discussed in chapter 3.3.1

The resulting models achieve average scan-match distances not significantly differ-

ent from bi-gram based models (see table 5.2). However, the distribution of tri-grams is entirely different compared to those produced by bi-gram based model. While for the bi-gram based model (see figure 5.3) the tri-grams real and simulated occur with different frequencies, for the model incorporating shallow parsing of the which differs much less (see figure 5.4).

It can be noticed that the achieved distribution match is better on the dataset with modulated saliency. This effect can be explained by more consistent subjects' behaviour resulting from lesser ambiguity in the stimuli (or rather emphasizing one of the stimuli's interpretations).

Overall performance is therefore much better, as not only the short, two fixation sequences match the real human behaviour, but also longer, more complex sequences that were not modelled accurately by bi-gram based model match human behaviour.

5.3 Summary

The presented series of models introduce a general framework for analysis and prediction of scan-paths in multi-modal visual and linguistic processing. The models differ from the existing work in several important aspects.

For example, recent work of Kukona and Tabor (2011) focuses on prediction of attentional shifts only at certain, discrete time points - reflecting word onsets. However, the eye movements do not happen at such fixed intervals - possibly more than one saccade can occur during each modelled time step. Our sampling based models are capable of generating such behaviour.

Moreover in contrast to work of Kukona and Tabor (2011) we train our models on experimental data, rather than rely on hand-wired model parameters.

Finally, we work with stimuli that has complex linguistic and visual stimuli components, unmatched by any recent work.

The presented results show that scan-paths carry interesting information that is neglected due to the popular belief that the scan-paths are not consistent across humans. We show that not only it is possible to study scan-paths at levels going beyond simple statistics such as proportion of looks at certain time points, but also to build models capable of simulating human behaviour in the presence of linguistic material as in investigated datasets.

We also show that linguistic material, particularly its current interpretation affects the attention, and scan-patterns (e.g. compare results of *Sampling (Symbol Phrases)*

with alignment methods in table 5.2).

The presented work also reveals the possibility of using information about the higher level structure of scan-paths, and applicability of tools and methodologies used in other fields, such as language processing, to the problem of predicting where people look while processing experimental stimuli in structured tasks in order to improve predicted behaviour.

Through this chapter we have neglected the influence of visual features on attentional selection. Even though we do not entirely discard possibility of saliency being important, we have already discussed its inappropriateness as a component of Visual World Paradigm data models. We investigate alternative approaches in chapter 7 along with a formulation of unified models.

Chapter 6

High level structure of scan-paths

6.1 Introduction

The main results presented by Visual World studies show both global and local correspondence of fixations to speech. For example, the classic study of Tanenhaus et al. (1995) discusses fixations falling on the objects just mentioned in the linguistic input. Coco (2011) shows local changes to the distributions of fixations upon disambiguation and results of studies such as Coco and Keller (2009) show that the dynamics of the scan paths go beyond simple bi-gram relationships. For example, within an NP the amount of fixations on the target increases after the onset of the noun. A simple n-gram model is not sufficient for modelling such dependencies, suggesting that scan paths might have some higher level structure.

These findings, and the possibility of generating human-like scan-paths as described in previous section, suggest that it is possible to learn and extract knowledge about higher level scan-path structure - going beyond quantitative analysis of collective behaviour or analysis of 'bi-grams' of fixations. We believe this topic deserves further investigation before we continue the discussion of eye-movement prediction.

6.2 Models

As already discussed, the natural extension to the model presented in section 5.2.3.1 would be to use longer n-gram language models. This approach poses some serious problems related to amount and sparseness of the available data. Assuming only five visual objects we have 25 (allowing re-fixations) bi-grams possible. However, the distribution of their probability changes along the course of the sentence, and with 14

Tag	Meaning
B	beginning of a fixations group (called chunk)
E	end of a chunk
I	fixation within a chunk
O	fixation not belonging to any chunk

Table 6.1: The list of tags used by the shallow parsers described in this chapter.

possible sentence partial parses (as in the RWP data used in this study) the number of distinct bi-grams grows to 350. For tri-gram these numbers are 125 and 1750 respectively. Considering the amount of scan-paths obtained in a typical eye-tracking experiment (i.e. 1500-2500) this approach quickly becomes infeasible.

An alternative approach to the problem is an analysis of sub-sequences of fixations in a manner similar to syntactic parsing or chunking (Abney, 1992). Thus far the models presented have exploited the syntactic structure of linguistic input. We will now focus on investigating and exploiting possible groupings of fixations. We formulate our approach as a tagging task similar to the work of Molina and Pla (2002) and Sha and Pereira (2003). We specify tags and define models such that tagging results reflect partial bracketing of the sequences, in manner similar to *syntactic chunking* of linguistic input (i.e. sentences).

Each fixation is labelled with one of the tags listed in Table 6.1 The constraints on the structure of chunks (i.e. chunks always starts with B, and ends with E) are ensured by forcing transition probabilities that violate this formulation to 0.

An alternative tag set, consisting only of I, B, and O tags, is also commonly used in the field of Natural Language Processing. However Pate and Goldwater (2011) presents discussion and evidence of the benefits of using additional E tag, as it allows forcing the chunks to be at least two fixations long (i.e. $B + E$ chunk). Moreover as we aim towards simultaneous modelling of scanpaths and sentences, we expect the boundary conditions to be worth explicit modelling, as it is the case for syntactic chunking in NLP.

We implemented the syntactic chunkers as Hidden Markov Model based taggers. HMMs are proven to be effective in supervised and unsupervised tagging and classification in various applications including, but not limited to part-of-speech tagging, speech recognition, and most importantly the shallow parsing itself. At the same

time HMMs are well understood, relatively simple, and a widely accessible technique, which can be applied to both supervised and unsupervised learning tasks. The architectures of the discussed models are presented in figure 6.1, and can be contrasted with models shown in figure 5.2 - where the first model is a simple HMM (as used in e.g. speech recognition) which treats the input S (i.e. sentence) as the observed stream, while trying to recover stream F, the underlying states encoding the latent stream (here scan-path). This type of model was discussed in section 5.2.2.1. The second model is similar, however the states do not encode fixations, but rather probability distributions. These are used to sample scan-paths using Markov-chain methods. This approach was presented in section 5.2.3

In this section, we will assume both sentence and scan-path to be observable. Instead of recovering the fixated objects, we will focus on recovering underlying groupings of fixated objects. The key concept is that the latent stream does not represent a sentence, but rather higher level groupings (i.e. chunks) of fixations. This can be seen as equivalent to syntactic chunking in language processing. The inputs in all cases are streams of discrete symbols representing sentences and scan-paths, exactly the same as used for the modelling in section 5.

The architecture of a basic model following this concept can be seen in figure 6.1 (top). This model reflects the description of an HMM tagger given in chapter 2.3.1.2. The probability of a parse can be expressed as the probability of underlying state sequence:

$$P(C,F) \quad = \quad a_{c(0)c(1)} \prod_{t=1}^{T-1} b_{c(t)}(f_t) a_{c(t)c(t+1)} \tag{6.1}$$

where $C = c(0),...,c(T)$ is the sequence of states (effectively chunk tags), $a_{c(t)c(t+1)}$ and $b_{c(t)}$ transition and emission probabilities, while $F = f_1,...,f_T$ is the scan-path being parsed.

The extension of this idea considers both the scan-path and sentence as observable sequences resulting from the same markov process. This can be represented as an HMM with multiple outputs as presented in figure 6.1 (top right). This approach can be seen as similar to treating pairs of symbols representing fixations and sentence parts as observable streams (instead of having 2 observable streams). The probability of the parse is given by:

$$P(C,S,F) \quad = \quad a_{c(0)c(1)} \prod_{t=1}^{T-1} b_{c(t)}(f_t) d_{c(t)}(s_t) a_{c(t)c(t+1)} \tag{6.2}$$

where again $C = c(0),...,c(T)$ is the sequence of states (effectively chunk tags), $a_{c(t)c(t+1)}$, $b_{c(t)}(f_t)$ and $d_{c(t)}(s_t)$ transition and emission probabilities, while $S = s_1,...,s_T$ and $F = f_1,...,f_T$ are the sentence and scan-path being parsed.

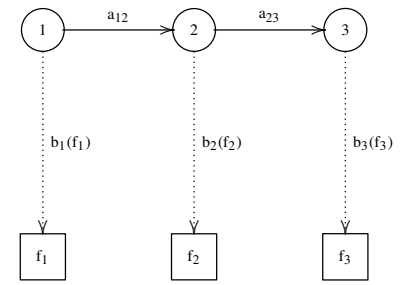
Finally, following work of Nefian et al. (2002a,b) and its application to syntactic chunking in Pate and Goldwater (2011), we consider scan-paths and sentences to be observable streams generated by coupled markov processes as presented in the bottom of Figure 6.1. Such model consists of two separate HMMs, however the transitions to next states in each of them depend not only on the current states, but also on the current states in the coupled HMM. The joint probability of the sequences and their labelling is given by:

$$\begin{aligned} P(C,S,F,X) &= a_{c(0)c(1)}e_{x(0)x(1)} \\ &\quad \cdot \prod_{t=1}^{T-1}\left[b_{c(t)}(f_t)a_{c(t)c(t+1)|x(t)}\right] \\ &\quad \cdot \prod_{t=1}^{T-1}\left[d_{x(t)}(s_t)e_{x(t)x(t+1)|c(t)}\right] \end{aligned} \tag{6.3}$$

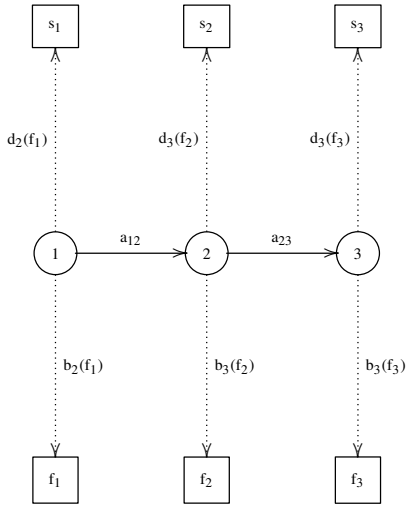
where again $C = c(0),...,c(T)$ is the sequence of states in the fixation HMM stream (effectively chunk tags), $X = x(0),...,x(T)$ is a sequence of states in the second (sentence) HMM stream, $a_{c(t)c(t+1)|x(t)}$, $b_{c(t)}(f_t)$ and $e_{x(t)x(t+1)|c(t)}$, $d_{c(t)}(s_t)$ transition and emission probabilities, while $S = s_1,...,s_T$ and $F = f_1,...,f_T$ are the sentence and scan-path being parsed. The stream X modelling a sentence does not have any particular interpretation, even though it can be seen as a chunking of a sentence. Its purpose is to increase the expressiveness of the model.

It is important to note that the transitions between states depend not only on the preceding state in the considered stream, but also the preceding state in the other - coupled - stream. Basic and Two-Output HMMs have only 4 states - reflecting each possible tag, and the constrains are enforced by ensuring zero-probability transitions where appropriate. The Coupled HMM has 4 states in each HMM path.

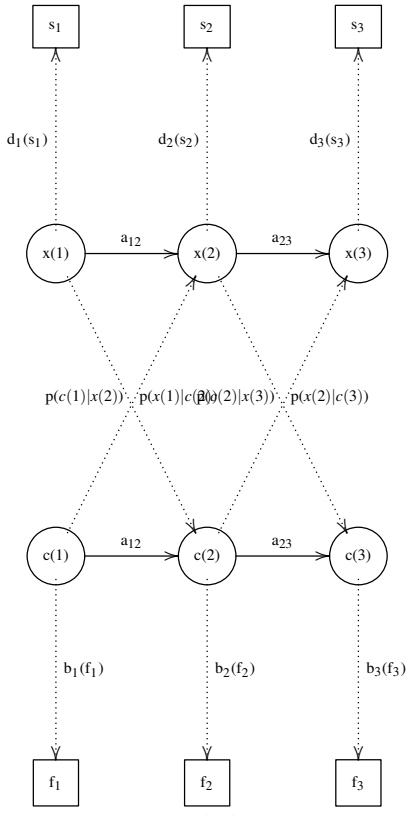
The training of the Basic HMM parser is achieved through the Baum-Welch procedure. It is possible to extend this algorithm for multiple emission probabilities of Two-Output HMM. The training of Coupled-HMM deserves more attention. As in other cases the training follows the general Expectation-Maximization procedure, with the junction tree algorithm (Pearl, 1982) applied in the E-step. Junction tree algorithm can be seen as a generalisation of a common Forward-Backward procedure to graphs. The exact procedure is out of scope of this thesis, and can be found in Liu et al. (e.g. 2002); Nefian et al. (e.g. 2002b).



(a) Basic HMM



(b) Two-Output HMM



(c) Coupled HMM

Figure 6.1: HMM architectures used in high level analysis of scanpaths. From top to bottom - Basic HMM, considering only the scanpath, Two-Output HMM modelling both scanpath and sentence sequences using the same state transitions, and Coupled HMM modelling scanpath and sentence with separate, yet coupled HMMs.

6.3 Evaluation

The evaluation of the models and the concept of chunking of scan-paths is a non-trivial issue. Typically, chunking and parsing of linguistic material are evaluated by a comparison to gold standard parses. However this approach is not possible here as the notion of scan-path syntax is non-existent, thus gold parses cannot be produced.

As a result the models are evaluated in a very simple manner by calculating coverage of the test data by chunking learnt on a training set. Additionally we apply the models to a higher level task not directly related to chunking, but requiring a deep understanding of the global and local structure of scan-paths - classification of experimental subjects. This task relies on the ability of the models to learn local, individual, and possibly very subtle differences in behaviour between people.

For the coverage calculation we used 10 fold cross validation. The data was split randomly, ensuring samples for each subject and trial are present in training set. The trained models were applied to remaining part of the data. The results are summarised in table 6.2. It is clearly visible that a very simple HMM can account for most of the unseen data. The results were analysed with an ANOVA to test for statistical significance by performing pairwise statistical tests between the compared models.

Not surprisingly the coupled HMM performs better than the simpler basic HMM ($F(1,84) = 7.36, p = 0.014$). A little surprising might be the fact that the basic HMM is significantly ($F(1,84) = 176.3, p < 0.001$) better than two output HMM. However it is important to remember that even though the two output HMM seems to be more powerful model it is essentially equivalent to a basic HMM with more complex observable sequence, and as such requires more training data (which is severely limited in this study).

The analysis of the lengths of chunks discovered on previously unseen data reveals a very interesting result - the majority of the chunks found are longer than 2 fixations,

Model	Data coverage (%)
Basic HMM	78.62 ± 2.69
Two Output HMM	66.17 ± 1.25
Coupled HMM	82.93 ± 4.24

Table 6.2: Coverage of previously unseen data with chunks found by various models.

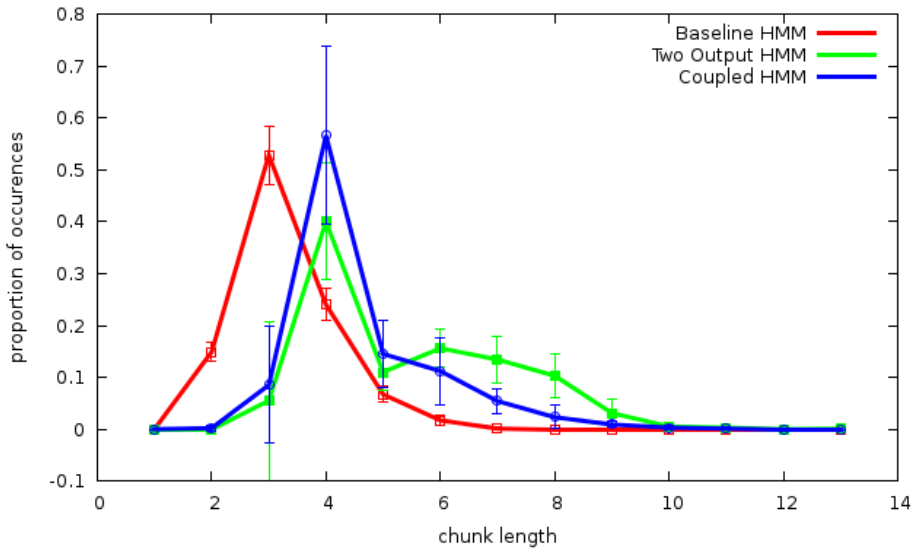


Figure 6.2: Histogram of lengths of chunks found on previously unseen data. Values averaged over 10 fold cross validation

supporting the initial hypothesis of scan-path structure going beyond simple bi-gram structure. The averaged histogram of chunks lengths can be seen in figure 6.2.

The second part of evaluation is based on a practical task: classification of subjects. This was also performed with 10 fold cross validation, with the eye tracking data of each subject being split among training and test sets. The set of HMMs is trained for each subject using corresponding samples from the training set. Test samples are then passed through this bank of HMMs resulting in log-likelihood of a sequence for each of the subjects. The sample is assumed to be produced by a subject whose model explains it best in terms of probability.

As a first baseline we used classifiers constructed using a modified version of scan-match (referred to as Needleman-Wunsch in table 6.3) - a distance of a tested sample to each of the samples in the training set is calculated. Then the tested sample is assumed to be produced by the subject who minimizes the following functions:

$$\hat{subject} = argmin_{subject} argmin_i d(x, s_i^{subject})$$

or

$$\hat{subject} = argmin_{subject} \frac{1}{N_{subject}} \sum_i d(x, s_i^{subject})$$

where $d(x, s_i^{subject})$ denotes distance between tested sample x and known i-th sample $s_i^{subject}$ produced by $subject$.

Model	Accuracy (%)
Chance level	4.00
Min Needleman-Wunsch	7.10 ± 4.81
Avg Needleman-Wunsch	8.81 ± 3.00
bi-gram log-probability	2.29 ± 0.17
kNN Needleman-Wunsch (best)	8.62 ± 3.96
Basic HMM	21.75 ± 3.04
Two Output HMM	21.56 ± 3.25
Coupled HMM	21.22 ± 3.50

Table 6.3: Classification performance of various models.

Additionally we include the kNN classifier into this set of baselines. The kNN classifiers are constructed from the regular minimum distance classifier, by selecting k training samples with lowest distance to considered sequence. The subjects proposed by these are considered to be votes toward final solution. In case of a tie, the subject is selected by applying method described by equation 6.4 to the voting samples involved in the tie.

The second baseline uses models developed in previous chapters. Transition probability distributions are learnt from the training data for each subject. The log-likelihood of each test sample is calculated under probability distributions corresponding to each subject, and the subject with highest probability is chosen.

This approach suffers from the problem discussed in section 6.2 - the inability to accurately estimate large numbers of probability distributions from the collected data. In this experiment only 16 to 18 training samples per subject are available, hence the estimated probabilities need heavy smoothing. For the sake of this experiment we used simple add-one smoothing.

In addition it is important to note that training samples come from several different experimental conditions (i.e. different sentence structure, saliency modulation), contributing to overall difficulty of the training.

Table 6.3 presents summarised results of the experiment performed on VAVWP dataset. It is clearly visible that classifiers based on either of the HMM chunkers outperform all of the baselines. The different chunkers do not differ significantly from

each other. They are clearly much better than all the baseline classifiers - analysing the results using ANOVA reveals significance in all cases, e.g. $F(1,84) = 91.73, p < 0.001$ for average Needleman-Wunsch compared with Basic HMM.

It is also important to notice that classifier based on techniques developed in previous section - bi-gram log-probability - performs worse than the chance level, and significantly worse than any distance-based baseline ($F(1,84) = 35.12, p < 0.001$ comparing with Avg Needleman-Wunsch). This can be explained by a lack of the training data needed to estimate a language model used in calculation of sequence probability - as discussed above in section 6.2.

The final observation is that all of the chunking HMMs perform classification at the same level of accuracy - it is particularly surprising is to see that the Basic HMM is not worse than the more advanced methods. This can be explained by the shape of the experimental data used in the evaluation - as already presented in section 3.3.1 all trials are constructed with only 3 types of sentences. Moreover the sentences are the same across the subjects (see section 5.2.3 for details). Nonetheless linguistic information helps the models to achieve better fit (coverage) on test data, and most likely is necessary to obtain similar results on more complicated datasets.

6.4 Parsing-Based scan-path comparison

6.4.1 Introduction

As mentioned in the previous chapters, a comparison of scan-paths is a non-trivial issue with no widely accepted solution. This issue is currently addressed by multiple studies resulting in an emergence of similarity metrics, computational toolboxes, and examples of their applications.

The most popular scan-path comparison methods are briefly discussed below highlighting their strengths and weaknesses.

Visual Recurrence is the simplest technique for comparing ordered sequences. This method relies on calculating an overlap between two sequences. As such it can only deal with sequences of an equal length.

Number of related techniques that can handle slight temporal shifts and non-linearities exist such as *cross recurrence analysis*, which has been successfully applied to geological temporal data (Marwan and Kurths, 2002), for alignment of mother

and child speech in linguistics (Dale and Spivey, 2006), and vision-language research (Richardson et al., 2007).

Edit-distance often referred to as *Levenstein distance*. In this method, the sequences of fixation points are represented as strings. Each of the string symbols represent a particular location (e.g. an object or an image patch) that is fixated. A globally optimal alignment between sequences is computed, associated with a score interpreted as a number of transformations (insertions, deletions, substitutions) required to transform one string into another.

Although easy to compute, this method suffers from several problems. Firstly it does not incorporate the notion of time other than the strings being an ordered sequence of symbols. One solution to this problem might be dividing the time-course into sequence of small time windows and have each of them represented by a separate string symbol.

The second problem is related to the edit-distance metric design - computation of the globally optimum solution. As a result, the score does not take existing local similarities into account. Figure 6.3 presents an example of such situation, where two of the presented sequences are clearly more similar to the template (first from left) than the remaining one.

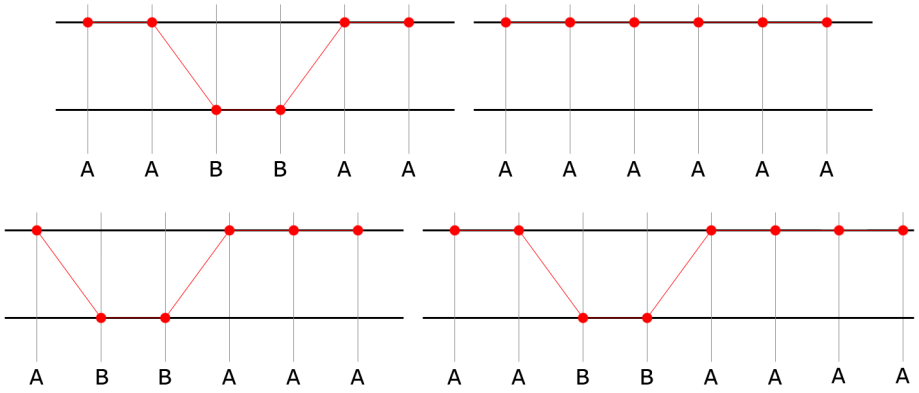


Figure 6.3: The template (top left) needs at least two operations to be converted to either of the three remaining sequences, regardless of its clear similarity to the sequences in the bottom line.

Ordered Sequence Similarity proposed by Gomez and Valls (2009), considers sequences of symbols of objects fixated during time periods of equal length. In its first

step, it identifies the objects occurring in both input sequences. Subsequently, it calculates a distance between corresponding objects. The distance is calculated as number of time periods between occurrences of the object in both streams normalised by the total length of the sequences.

Although OSS was shown to be more effective than the edit-distance, it still suffers from inability to handle small differences in the onset and length of fixations (see figure 6.4).

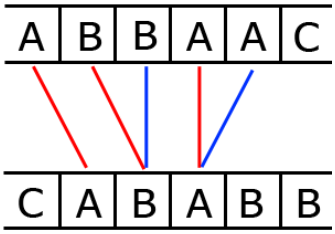


Figure 6.4: The OSS considers only concurrencies of the fixated objects (red lines). However it misses obvious relations between sequences (blue lines)

Alignment methods such as Needleman-Wunsch or Smith-Waterman algorithm have also been applied to compare scan-paths. They are not only capable of handling sequences of different length, but also can compensate non-linearly for small timing differences. This class of methods is used in the majority of currently available toolboxes such as ScanMatch (Cristino et al., 2010).

In principle they find optimal alignment between 2 sequences, although usually the minimal distance computed as the side effect is the subject of interest. The computed alignment is usually only globally optimal, and most of the alignment algorithms can be interpreted as generalizations of the Levenstein distance.

Through this thesis, we have used ScanMatch as it is the technique that is the widely accepted and used in analysis of eye-tracking data not only in vision and language but many other fields such as studies of eye-hand coordination (Leek et al., 2012), problem solving (Madsen et al., 2012).

Although ScanMatch does not address all the issues of the techniques mentioned above (e.g. it does not have integral notion of time and duration) its architecture allows to providing ad-hoc solutions where necessary being very simple and computationally inexpensive. Moreover due to its recent popularity it allows direct comparison with results reported by other studies.

Multimatch (Jarodzka et al., 2010) is a vector based, multidimensional approach to comparison of scanpaths. Instead of relying on one similarity score, it integrates collection of techniques that consider shape, order, direction, timing and several other characteristics of eye movements.

Each of the components addresses one or more of the weaknesses of other techniques. Moreover it is open to addition of further components offering a great platform for comparing any type of spatio-temporal sequences.

However despite being shown to offer great sensitivity to similarities (see e.g. Foulsham et al., 2012; Dewhurst et al., 2012) it did not gain popularity due to its relatively high complexity.

6.4.2 Comparing scan-paths with shallow-parsing

The results of applying shallow-parsers to the scan-paths presented in chapter 6 shows that it is possible to learn a general, higher level structure of the fixation sequences. At the same time all the popular methods of scan-path comparison are based on direct alignment of the fixated points or areas rather than patterns of behaviour. This can be seen as an equivalent of comparing natural language sentences word-by-word instead of comparing their syntactic structure.

The possibility of learning the high-level structure of scan-paths enables us to change the approach we use for measuring the similarity of scan-paths. In this section we will discuss some possible solutions based on the direct use of knowledge extracted with the HMM shallow parsers introduced in this chapter.

6.4.2.1 Sequence probabilities

The first and the most obvious method of comparing scan-paths is though a direct application of trained parsers to calculate sequence probabilities. This method, although very simple, requires a considerable amount of training samples to build an appropriate parser. Hence it is applicable only to tasks that require checking if a considered scan-path belongs to a predefined group (e.g. such as speaker identification discussed in section 6.3) and can not be directly applied to calculate a similarity between two scan-paths directly.

To cope with the task of measuring similarity between two scan-paths the construction of the HMM chunker has to be based on only one training sample (i.e. the

scan-path we want to compare the other scan-path to). Moreover the calculated similarity d should be symmetric, that is $d(S_1,S_2) = d(S_2,S_1)$.

The simplest approach to achieve this is by building parsers for each of the considered scan-paths. In this setting, the HMMs are trained using only one sample sequence. Due to a small amount of training data, it is necessary to ensure that all of the allowed transitions have non-zero probabilities, such that the HMMs can represent any valid scan-path sequence. This can be achieved by applying an appropriate smoothing.

Using the trained chunkers, we can compute the probability of the scan-paths they were trained on and compare it to the probability of the scan-path under the second model. Multiple methods of comparing these probabilities and combining them into one value representing the similarity between scan-paths are possible.

An obvious solution would be to use average probability of the scan-path under model corresponding to the second scan-path:

$$d(S_1,S_2) = \frac{p_{S_1}(S_2) + p_{S_2}(S_1)}{2} \tag{6.4}$$

where S_1 and S_2 are the compared scan-paths, and $p_{S_j}(S_i)$ is a probability of a parse of scan-path S_i produced by HMM parser trained with scan-path S_j.

The main issue with this method is that probability $p_{S_i}(S_i)$ of a parse is lower than 1 even for a model corresponding to parsed scan-path. This is because the models contain only small number of states necessary to perform tagging - increasing the number of states would result in better fit to the data, however would reduce models ability to generalize to unseen sequences. As a result it is more appropriate to either consider difference between probabilities under both constructed HMMs:

$$d(S_1,S_2) = \frac{|p_{S_1}(S_1) - p_{S_2}(S_1)| + |p_{S_2}(S_2) - p_{S_1}(S_2)|}{2} \tag{6.5}$$

or a geometric mean of their ratios

$$d(S_1,S_2) = \sqrt{\frac{p_{S_1}(S_1)}{p_{S_2}(S_1)} \cdot \frac{p_{S_2}(S_2)}{p_{S_1}(S_2)}} \tag{6.6}$$

6.4.2.2 Parser's HMM structure

We will also investigate an alternative approach, in which the constructed parsers are compared directly. Such an approach has been applied in the bioinformatics to compare genome sequences, in software engineering to study software behaviour, and in many other domains (see e.g. Soding, 2005; Lyngso et al., 1999).

In this particular case the general structure of the Finite State Machines representing HMM parsers is not changing. Therefore the differences in structure arise exclusively from their parametrisation - the transition and emission probabilities.

The simplest way of comparing the FSM is hence to analyse the differences by calculating average Kullback-Leibler divergence between corresponding transition and emission probability distributions:

$$d(S_1,S_2) = \frac{1}{N}\sum_{i=1}^{N} KL(b_i(S_1),b_i(S_2)) + \frac{1}{N}\sum_{i=1}^{N} KL(a_i(S_1),a_i(S_2)) \quad (6.7)$$

where N is number of states, $a_i(S)$ and $b_i(S)$ are a distribution of transition and emission probabilities from state i of HMM trained on sequence S.

6.4.3 Experiments and Discussion

We verify the discussed approach by applying it to an example discussed by Cristino et al. (2010). It consists of a simple task of fixating numbers of a given colour scattered over a plane in a given order. Figure 6.5 presents the stimuli and example scanpaths.

We encode the presented scanpath by simply identifying the digit (i.e. value and color) associated with each fixation point. The association is realised by checking if fixation falls into a circle enclosing the digit. The fixations that are not targeting any digit are encoded as 'background'.

It is worth noting that the information about duration of fixations is not preserved in this encoding. Nonetheless, any realistic scan-path comparison method should be able to distinguish scan-paths collected for both experimental conditions (i.e.: fixating red or blue numbers in ascending or descending order) in this task.

Table 6.4 presents an example of log-distances between different scan-paths calculated using the method defined by equation 6.5. As expected, the approach is able to cope with this simple dataset correctly identifying the similarity between scan-paths collected for the same experimental task.

We also apply the constructed measures to the classification task presented in the chapter 6.3. The results are summarised in the table 6.5 and are complementary to those presented in table 6.3.

It is obvious that the simple methods of comparison based on the probabilities of the scan-path parses are no better than the traditional, alignment-based approach using Needleman-Wunsch algorithm.

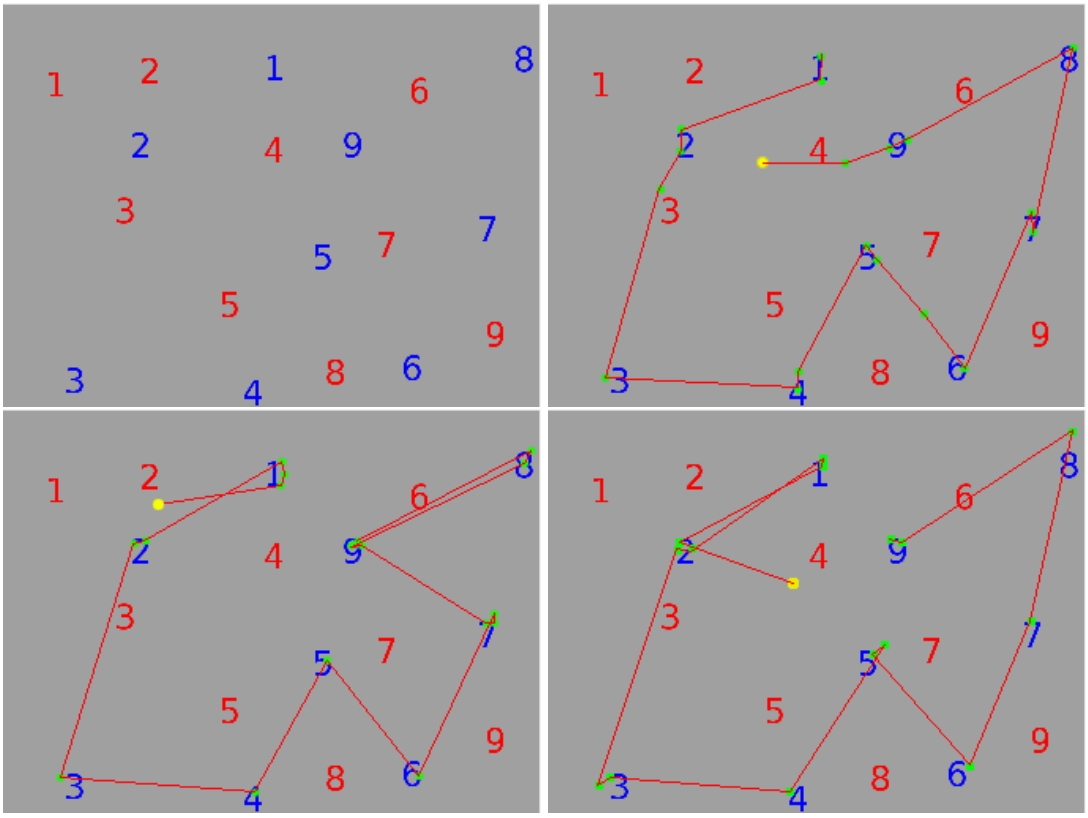


Figure 6.5: The stimuli and scanpaths collected for an initial evaluation of ScanMatch (Cristino et al., 2010). Top left: stimuli, top right: scanpath for fixating the blue digits in descending order, bottom line: two scanpaths for fixating the the in ascending order. The yellow dot indicates initial position of an eye gaze, green dots consecutive fixation points, and red line saccadic movement.

scanpath (condition)	1 (blue, asc)	2 (blue, asc)	3 (blue, desc)
2 (blue, asc)	-16.54	0	
3 (blue, desc)	-13.83	-13.82	0

Table 6.4: An example of log-distances between 3 sample scan-paths shown in figure 6.5 calculated using method defined by equation 6.5.

However it is important to note that these parsers are not built specifically with comparison of scan-path in mind. In addition, the training of the HMMs - based on Expectation-Maximization method - can have a serious impact on the performance of a parser over different runs. Nevertheless, the results presented in the tables 6.5 and 6.3 suggest possibility of using higher level structure of scan-paths as a basis for computing similarity - especially that in a practical task of classification, the parsers

Metric	Accuracy (%)
Chance level	4.00
Needleman-Wunsch (min)	7.10 ± 4.81
Parsing probability difference (min)	5.90 ± 0.03
Parsing probability ratio (min)	5.40 ± 0.03
Needleman-Wunsch (avg)	8.81 ± 3.00
Parsing probability difference (avg)	5.60 ± 0.03
Parsing probability ratio (avg)	7.45 ± 0.03

Table 6.5: Classification performance of various metrics.

were able to match Needleman-Wunch performance, and even beat it, provided enough samples to train the HMMs (see Basic, Two-Output an Coupled HMMs in table 6.3).

6.5 Summary

In this chapter we have discussed an indication of the existence of a high level structure in scan-paths during situated language comprehension. Statistical patterns of the eye-movements are observed in in the majority of the eye-tracking experiments. Experimental researchers commonly utilise statistical methods to investigate various phenomena in eye-tracking data. However, despite the hypothesis as early as Noton and Stark (1971), no structure in form of repeatable, coherent groupings of fixations had been definitively proven so far.

As a result, models of eye-movements do not include components controlling a syntactic structure of the scan-paths. Usually they do not consider possible dynamics of the movements beyond proportion of fixations falling on a given object.

The shallow parsing discussed in this chapter shows that it is possible to extract and build models of human behaviour at a level much greater than coarse analysis of global similarities. Chapter 5.2.3.2 gives an example of practical application of the parsers to the scan-paths prediction. The results indicate that controlling the structure of a scan-path by forcing predicted fixations to form larger groups is beneficial. The scan-paths predicted with such supervision of their syntactic structure resemble a real human behaviour much more than obtained with equivalent models based purely on

statistical correspondence between languages and eye-movements.

Our findings are fundamentally different from hypotheses of Noton and Stark (e.g. 1971); Cooper (e.g. 1974); Tanenhaus et al. (e.g. 1995): we do not assume a presence of fixation pattern that is repeated on subsequent viewings on the same stimuli. Rather than that, we show an emergence of shorter, structural groupings of fixations - chunks - that combined form a complete scan-path.

In addition to the observed syntactic structure, we have found an indication of consistent, subject-specific variations in behaviour. These individual characteristics can also be learnt and utilised. We show, that syntactic parsers trained to recognise individual differences in behaviour are able to outperform Scan-Match in subject-identification task, despite very small training set available.

In section 6.4 we investigate whether the syntactic structure of the scan-paths can be used for their comparison. This approach is entirely different from currently used methods that usually rely on alignment of considered scan-paths. The result obtained proof-of-concept implementation are very encouraging indicating sensitivity no worse than established methods such as Scan-Match.

Although global, statistical patterns of the collective human behaviour arise in most of experimental conditions, we believe that no other study has presented existence and successfully utilised the syntactic structure of the scan-paths. Moreover, we have shown that such structure is not only specific to experimental trial as it could be hypothesised in Visual World Paradigm setting (see e.g. Tanenhaus et al., 1995), but also to individuals.

The presented results suggest that analysis and modelling of the eye movements should take into account the structural constraints and subject-specific differences between them into account.

Chapter 7

Objectness and saliency

7.1 Introduction

Recent studies have shown that the allocation of visual attention appears to be primarily governed by factors beyond low-level salience, as presented by Itti et al. (e.g. 1998). Specifically results presented by Einhauser et al. (2008); Nuthmann and Henderson (2010) show there is greater correlation between fixation locations and the position of an object than with salient areas.

At the same time, practical applications of attentional models, e.g. in robotics, require object based selection of areas of interest (see Yu et al., 2010; Schauerte et al., 2012). Similarly, the problem of modelling synchronous processing studied throughout this thesis also requires the consideration of objects as a common interface for binding visual and linguistic processors.

In previous chapters we have investigated the visual attention using notion of *object*. In chapter 4 objects were used implicitly as a basis for computation of contextual bounds, although the model did not explicitly consider objects in later stages. The models presented in chapters 5 – 6.4, on the other hand, use the concept of objects as a fixated entity, throughout explicitly the whole modelling process.

We believe that the notion of objectness is crucial in order to build complete, integrated models of visual and linguistic processing. This chapter describes an investigation into one of the major problems associated with objectness - the correspondence between the visual features of an objects and the process of selecting it for visual attention. We hypothesize, that this low-level visual interestingness of the objects - which we will refer further to as *object saliency* - can influence perception at high, concious level. We investigate this hypothesis with a simple Mechanical Turk experiment and

by analysis of the Object Naming dataset described in chapter 3.4.1. Moreover we investigate whether models of high level tasks such as *Salient Object Detection* can predict human visual attention. We are not concerned with prediction of salient image patches, but rather with the selection of objects that are likely to be fixated. This approach allows us to develop computational models of attentional selection based on cognitive relevance defined over objects (Henderson et al., 2007, 2009).

7.2 Background

Sequences of fixations are important indicators of the processing performed by attentional systems and a number of models have been proposed to predict eye-movements during scene comprehension. They can be broadly divided into two categories. The first one consists of bottom-up models exploiting low-level visual features to predict areas likely to be fixated. A number of experimental studies have shown that certain features and their statistical unexpectedness attract human attention (Bruce and Tsotsos, 2006). The best-known example is Itti and Koch's model (Itti et al., 1998) which builds saliency maps based on color, orientation, and luminance filters inspired by neurobiological results. The second group of models assume the existence of top-down supervision of attention which contributes to the selection of fixation targets and a number of models have been proposed to capture context effects on visual attention. A prominent example is the Contextual Guidance Model (Torralba et al., 2006), which combines bottom-up saliency with a prior encoding global scene information.

At the other end of continuum there is the *cognitive relevance hypothesis* which holds that fixations are directed according to the requirements of the current task (Henderson et al., 2007). Although the attentional processing and fixation locations are generated from visual input they are assumed not to be ranked on basis of saliency, but rather based on their relevance to the current task. There is considerable experimental evidence showing, that saliency has only minor impact on fixation patterns (Henderson et al., 2009; Nuthmann and Henderson, 2010; Einhauser et al., 2008; Underwood and Foulsham, 2006). Pomplun (2006) shows, that effects of visual features depend on their relation to search target.

These two views – visual salience and cognitive relevance – differ in the representation over which attentional selection is made. Saliency requires low-level, bottom-up image representation, while cognitive relevance framework needs higher-level, object based representation. To the best of our best knowledge, there is no complete computa-

tional model based on the cognitive relevance hypothesis. The closest work is perhaps that of Spain and Perona (2011), who developed a model for object importance (defined as the probability of an object in a scene being named) which includes several features derived from saliency maps. Related work of Einhauser et al. (2008) shows that the location of objects in a scene is a better predictor of fixations than bottom-up, pixel based saliency.

It seems, however, that some objects will naturally attract more fixations than others. This intuition was discussed by Rensink (2000b,a) with a proposal of *proto-object* - pre-recognition entities that draw attention, with a matching model proposed by Walther and Koch (2006). Proto-object based models have been, to some extent, successfully applied in robotics to create attentional systems for virtual and physical agents (see Yu et al., 2010; Schauerte et al., 2012). These models are not truly object-based, and work in a manner similar to image segmentation. They divide an image into collection of blobs, that correspond to areas enclosed by curves of constant, high saliency values, which are not likely to correspond to real objects (see problem depicted in figure 7.3 discussed in section 7.4.1) Figure 7.1 shows an example of proto-objects extracted from an image using the method proposed by Walther and Koch (2006).

The work of Nuthmann and Henderson (2010) provides evidence of human attention being object, rather than proto-object based. This questions the validity of using the above mentioned models for predicting human fixations. Models of object based saliency, or importance, have been proposed in computer vision. For example work of Liu et al. (2011) and Klein and Frintrop (2012) focuses on detection of salient objects, using a ground truth based on human annotation. The models use Machine Learning techniques to learn which arrangements of computable visual features such as center-surround histograms, orientation, scale etc. are likely to be perceived as salient. However, the task they solve is purely engineering: detection of areas matching the pre-annotated training data, rather than offering an explanation of human behaviour. Details of these methods are summarised in section 7.4.2.

An important problem of off-object fixations arises naturally in context of object-based attentional selection. Several causes of such fixations can be identified. *Target Acquisition Model*(TAM) model of Zelinsky (2008) was shown to explain off-object fixations by objects population averaging. The averaging seems to occurs across the entire scene early in the scene viewing resulting in more central fixations, while smaller populations are considered later with fixations falling inside the groups rather than

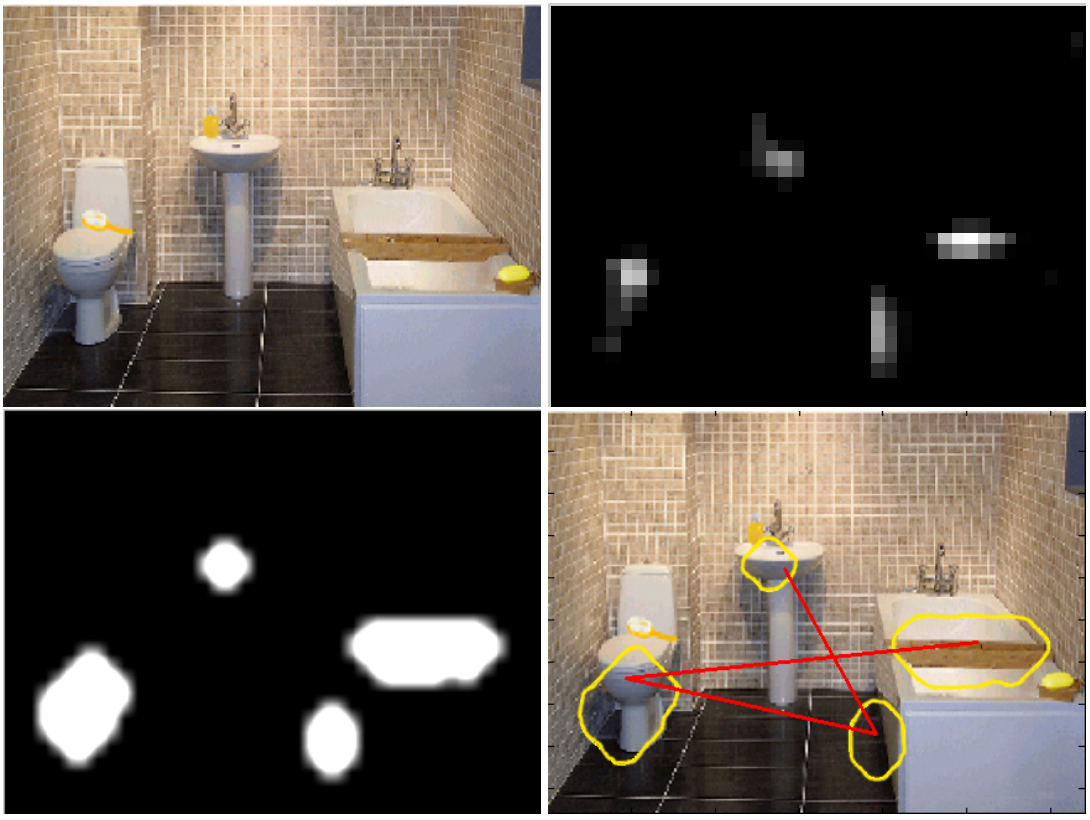


Figure 7.1: Example of proto-objects extracted from an image using Walther and Koch (2006) toolbox. From left to right, top to bottom: original image, saliency map computed according to Itti et al. (1998) method, proto-object mask, and finally an image showing proto-object and scan-path simulated by a winner-take-all network. It can be seen, that the salient patches, and consequently proto-objects do not necessarily correspond to the real objects in the scene.

specific objects (Zelinsky, 2012). In addition the study presented by Nuthmann and Henderson (2010) indicate possibility of some of the off-objects fixations being a result of oculomotor system inaccuracies - especially in case of longer saccades to small objects. It is also important to remember that eye-tracking software relies on certain - often configurable - criteria in order to differentiate saccades and fixations. Fixation detection and eye pupil position determination might not be error-free resulting in false fixations being recorded.

The work of Spain and Perona (2011) investigates the likelihood of an object being mentioned in a naming experiment. They refer to this as *object importance*, and develop a model based on linear combination of object features such as area, position and saliency, that is capable of predicting object's importance. Moreover it shows, that

it is not possible to extract important objects using the state of the rate segmentation methods.

Our attempt to answer the question of how the objects are selected by attentional systems will be limited to the investigation of the visual processing path. The studies mentioned above indicate, that there is a preference for the fixations to be targeted towards object centres, rather than salient areas. In section 7.3 we show, that not just any object centre is selected by attentional mechanisms, even in tasks that are traditionally assumed to have very little top-down guidance. This result, in a connection with existing experimental evidence of attentional selection being influenced by visual features, suggests that the visual attention might be attracted to saliency on per-object basis, rather than per area.

This assumption brings an interesting problem of calculating object visual interestingness. Several methods proposed in the literature, or developed by us are summarised below.

7.3 Preferred Landing Position

Nuthmann and Henderson (2010) argues that the majority of fixations are directed towards objects. Moreover the fixation locations seem to be normally distributed around centre of objects. This raises a natural question whether low level saliency and object based selection (or rather preferred landing position) can be combined into one coherent framework.

Intuitively, saliency and preferred landing positions might be closely related and experimental results could be explained by a simple mechanism, in which the selection of areas of interest is influenced by visual features, while the fixations themselves are directed towards centres of the objects in the selected areas.

We begin to investigate this hypothesis by building and evaluating a simple model that extends saliency with information about object centres. In our model the final map M at position $p = (x, y)$ is computed as sum of regular saliency map S and overlay C carrying information about objects positions and sizes:

$$M(p) \quad = \quad S(p) + C(p) \tag{7.1}$$

We compute saliency map S following Torralba et al. (2006) (see chapter 4.3.1.1 for details). The overlay C is computed as a sum of two-dimensional Gaussians centred

at each object, with spread proportional to the object size:

$$C(p) = \sum_{o \in O} \mathcal{N}(p, \mu_o, \Sigma_o) \tag{7.2}$$

where O is the set of annotated objects in the image, μ_o and Σ_o are parameters of the Gaussian associated with an object o. μ_o is defined as the object's centre of mass. Σ_o on the other hand is a diagonal covariance matrix of following structure:

$$\Sigma_o = \begin{vmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{vmatrix}$$

with $\sigma = \tau\sqrt{a^2 + b^2}$, where a, and b are dimensions of the objects bounding box, while τ is a normalization constant to be chosen experimentally.

This formulation puts emphasis on the centre of the object, bounding it at the same time to its size. As a result, small objects are perceived as more interesting, than large, often sparse areas such as the sky.

7.3.1 Evaluation

We evaluate the model against regular saliency, by calculating the proportion of fixations falling on areas of image with highest values of saliency. The details of this procedure can be found in chapter 4.3.2. We also investigate the effect of using the centre overlay alone without modulating by the saliency map.

The models are evaluated over two datasets: the object counting dataset described in chapter 3.3.1, and object naming dataset described in chapter 3.4.1. These datasets are chosen as they are provided with satisfactory object annotations. Furthermore, they represent two different tasks requiring different level of contextual processing.

The results are summarised in terms of receiver-operating characteristic in figure 7.2. It is immediately visible, that modulated saliency and object overlay is superior to saliency for thresholds smaller than 40%. This confirms, that people have preference to fixate object centres rather than arbitrary salient areas.

We test the results for statistical significance by performing pairwise ANOVA between results obtained with evaluated models. The analysis of the area under ROC curves summarised in table 7.1 reveals, that for both datasets, the object position overlay is significantly better than saliency with $F(1,24) = 9.27, p < 0.005$ for object counting, and $F(1,23) = 9.84, p < 0.005$ for object naming.

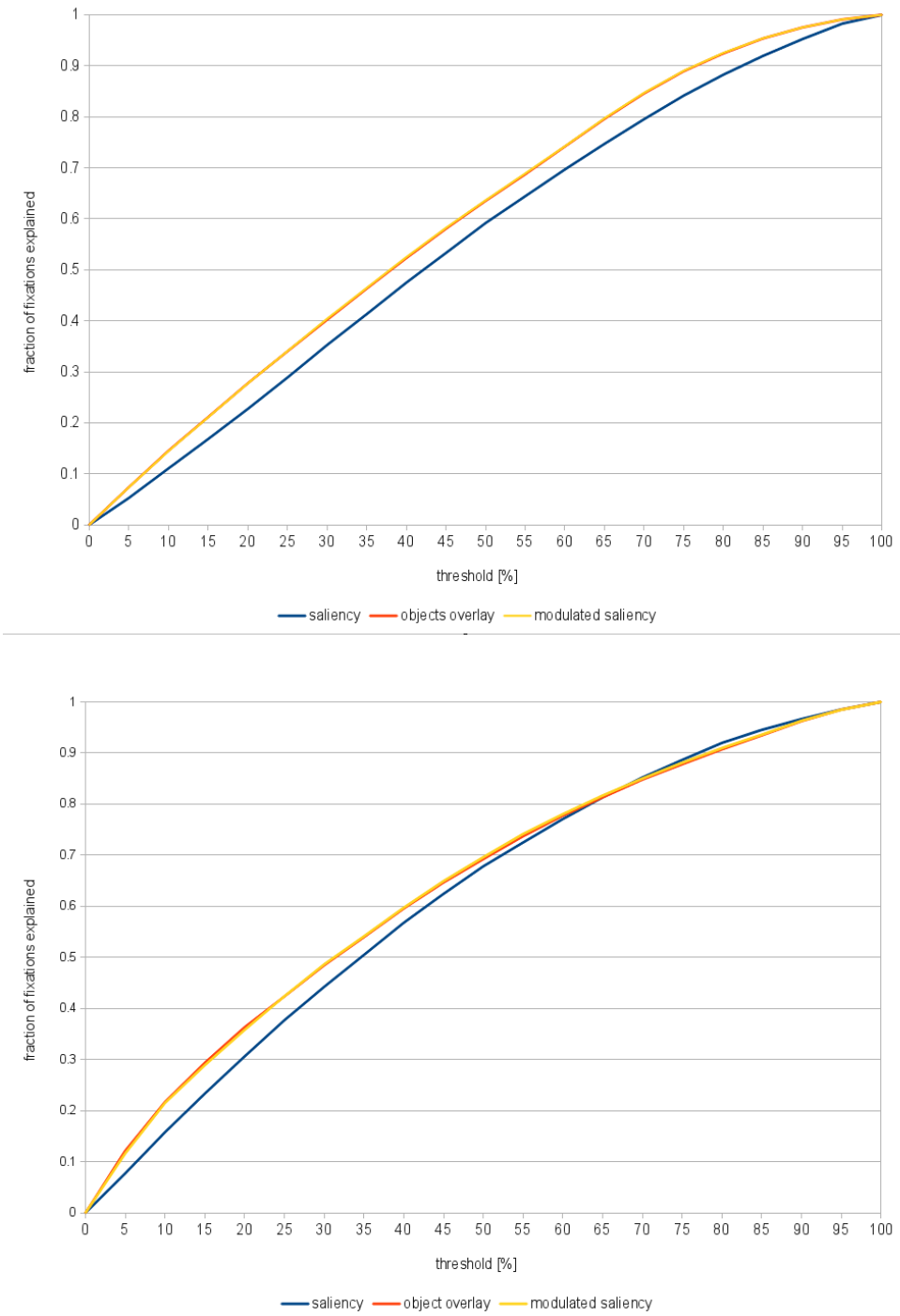


Figure 7.2: Effects of modulating saliency with object positions overlay map in object naming task (top), and object counting task (bottom)

Combined model using both object positions and saliency does not improve the result further - the improvement over saliency is significant ($F(1,24) = 11.12, p < 0.005$ for counting, and $F(1,23) = 10.13, p < 0.005$ for naming), but no different than

Model	Object counting	Object naming
Saliency	61.66%	55.87%
Object overlay	63.60%	59.78%
Modulated saliency	63.67%	59.85%

Table 7.1: Results of predicting fixated areas based on notion of *Preferred Landing Position* as area under ROC curves.

object position overlay alone ($F(1,24) = 0.08, p = 0.78$ for counting, and $F(1,23) = 0.01, p = 0.93$ for naming).

This improvement is however very small, suggesting that not just any object is fixated in its center, but only some of them. This leaves an important question which objects are selected, and how the visual features contribute to this mechanism.

7.4 Calculating object-based saliency

7.4.1 Conversion of standard saliency

Given the availability of various saliency models, a natural approach is to convert the results produced by such models into object based representation.

Several methods of such conversion are possible, including, but not limited to the calculation of simple statistical measures such as the mean or maximum of the saliency values over all the points belonging to the object:

$$OS_{max}(o) = \max_{(x,y)\in o} S(x,y) \tag{7.3}$$

$$OS_{mean}(o) = \frac{1}{N}\sum_{(x,y)\in o} S(x,y) \tag{7.4}$$

where $OS(o)$ is an interestingness value of objects o, S is a saliency map, and x, y are coordinates of the points in the image.

We also investigate median and mode of the saliency S values within the object. These methods are referred to as *converted* in figure 7.5 and table 7.3.

This solution is however ad-hoc, and we suspect it is a subject to an anomaly present in various saliency models: the highest saliency values area associated with

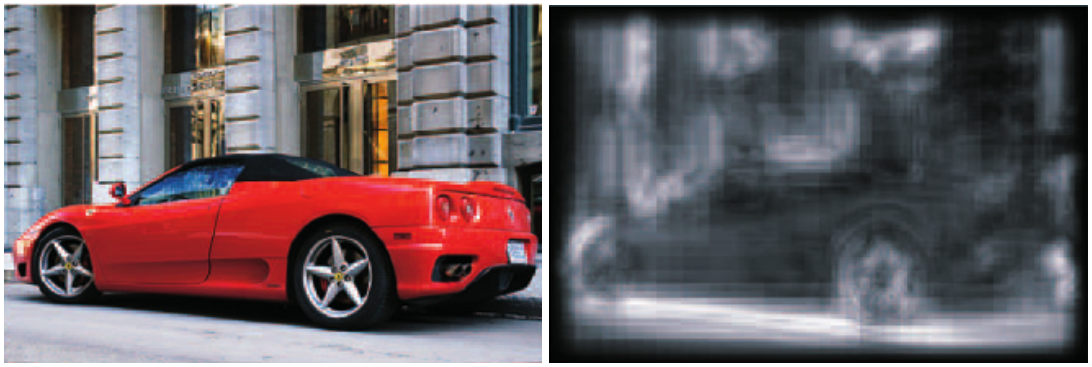


Figure 7.3: An example of anomaly often occurring while calculating pixel-based saliency maps adopted from Liu et al. (2011). The highest saliency values are associated with sharp edges often being the boundaries of the objects, while larger, more constant areas inside the objects are not salient according to the model.

edges of the objects. We believe this anomaly might lead to the situation where the saliency associated with the object is scattered over its neighbourhood (see figure 7.3 for example).

7.4.2 Liu et al. 2011 salient object detection

In order to overcome problem mentioned above, Liu et al. (2011) proposes an alternative approach to quantifying visual appearance. The proposed solution formulates the detection of salient objects as a binary labelling problem, where pixels are tagged whether they belong to a salient object or not.

The problem is approached using a Conditional Random Field (CRF, Lafferty et al., 2000) framework. The probability of labelling if configuration of pixels belong to a salient object, is given as:

$$P(A|I) = \frac{1}{Z}\exp^{-E(A|I)} \tag{7.5}$$

where $A = a_x$ is a set of labels assigned to pixels, $a_x \in 0,1$ is a binary label indicating whether pixel x belongs to a salient object, Z is a normalization constant (partition function), and energy $E(A|I)$ is defined as weighted sum of K unary salient features $F_x(a_x, I)$, and a pairwise feature $S(a_x, a_{x'}, I)$:

$$E(A|I) = \sum_x \sum_{k=1}^{K} \alpha_k F_k(a_x, I) + \sum_{x,x'} S(a_x, a_{x'}, I) \tag{7.6}$$

with α_k is a weight of k-th feature, and x, x' are adjacent pixels.

The main advantage of CRF is that the features can be arbitrarily low or high-level, resulting in elegant and optimal framework to combine multiple features.

The feature $F_k(a_x, I)$ is a binary flag, indicating whether, based on corresponding visual feature, the pixel x belongs to a salient object:

$$F_k(a_x, I) = \begin{cases} f_k(x, I) & a_x = 0 \\ 1 - f_k(x, I) & a_x = 1 \end{cases} \tag{7.7}$$

where $f_k(x, I) \in [0, 1]$ is a normalized map of features for every pixel x.

The pairwise feature $S(a_x, a_{x'}, I)$ is defined using the contrast-sensitive potential function of Boykov and Jolly (2001):

$$S(a_x, a_{x'}, I) \quad = \quad |a_x - a_{x'}| \cdot \exp^{-\beta d_{x,x'}} \tag{7.8}$$

where β is color contrast weight parameter Blake et al. (see 2004, for more details), and $d_{x,x'} = |I_x - I_{x'}|$ is a L2-norm of color difference at pixels x and x'.

$S(a_x, a_{x'}, I)$ can be interpreted as a penalty for assigning different binary labels to adjacent pixels.

The details of CRF, learning and inference are out of scope of this thesis and are discussed in details in Liu et al. (2011). Much more interesting is however selection of the visual features discussed in more detail below.

7.4.2.1 Multiscale contrast

Contrast is one of the most popular features used in saliency models. In this particular case, the contrast operator is defined as a linear sum of responses from Gaussian image pyramids:

$$f_c(x, I) = \sum_{l=1}^{L} \sum_{x' \in N(x)} |I^l(x) - I^l(x')|^2 \tag{7.9}$$

where I_l is the l-th level of the pyramid, L total number of pyramid scales (here $L = 6$), and $N(x)$ is 9x9 neighbourhood of pixel x.

The computed map $f_c(\cdot, I)$ is normalized into a fixed $[0, 1]$ range.

The contrast operator is meant to simulate human receptive fields, by highlighting the high-contrast boundaries, while omitting homogeneous regions inside objects.

7.4.2.2 Center-surround histograms

The typical approaches to saliency are based on detection of high-contrast center-surround areas. This however leads to the crucial problem mentioned earlier - the high scores assigned to boundaries of the objects.

To target this issue, Liu et al. (2011) proposes to use regional feature. The feature is defined as distance χ^2 between histograms of RGB color in the object and its surroundings:

$$\chi^2(R,R_S) = \frac{1}{2}\sum\frac{(R^i - R^i_S)^2}{R^i + R^i_S} \tag{7.10}$$

The histograms are computed over rectangular area R enclosing an object, and surrounding contour R_s, with the same area as R. The details of constructing these rectangles on unlabelled images can be found in Liu et al. (2011), and are not important for understanding the feature computation.

The feature function $f_h(x,I)$ is proportional to the χ^2 distance:

$$f_h(x,I) \propto \sum_{\{x'|x\in R(x')\}} w_{x,x'}\chi^2(R(x'),R_S(x')) \tag{7.11}$$

where $R(x')$ is a rectangle centred at pixel x' containing pixel x, and weight $w_{xx'} = \exp^{-\frac{|x-x'|^2}{2\sigma^2_{x'}}}$ with variance $\sigma^2_{x'}$ being one-third of the size of $R(x')$. The map $f_h(\cdot,I)$ is normalized to $[0,1]$ range.

7.4.2.3 Colour spatial distribution

The last feature used by Liu et al. (2011) is spatial color distribution. It is motivated by an observation, that salient objects are less likely to contain colors that are widely distributed through the image.

The simplest method to quantify this observation, is to compute spatial variance of color. The first step is to represent all colors by a mixture of Gaussians $w_c,\mu_c,\Sigma_{c\,c=1}^{C}$, where w_c,μ_c,Σ_c are the weight, mean and covariance of the c-th component, and C is number of components. Each pixel is then assigned the color component as:

$$p(c|I_x) = \frac{w_c\mathcal{N}(I_x|\mu_c,\Sigma_c)}{\sum_c w_c\mathcal{N}(I_x|\mu_c,\Sigma_c)} \tag{7.12}$$

The horizontal spatial variance is computed as:

$$V_h(c) = \frac{1}{|X|_c}\sum_x p(x|I_x)\cdot|x_h - M_h(c)|^2 \tag{7.13}$$

where

$$M_h(c) = \frac{1}{|X|_c}\sum_x p(x|I_x)\cdot x_h \tag{7.14}$$

while x_h is the x-coordinate of the pixel x, and $|X|_c = \sum_x p(c|I_x)$.

The vertical variance $V_v(C)$ is defined in the same way, with exception of considering y-coordinates.

The spacial variance $V(c)$ is normalized sum of both partial results:

$$V(c) = V_h(c) + V_v(c) \tag{7.15}$$

Finally, the feature function $f_s(x,I)$ is defined as:

$$f_s(x,I) \propto \sum_c p(c|I_x)\cdot(1-V(c)) \tag{7.16}$$

The map $f_s(\cdot,I)$ is normalized to $[0,1]$ range.

7.4.2.4 Applying salient object detection to interestingness score calculation

As we are interested in ordering of the objects according to their attractiveness, rather than in image segmentation, this model does not directly fit our purpose. However, it can be easily adopted to perform fixation prediction.

In our case the problem of segmenting the image (i.e. determining whether pixel belong to a salient object or not) is solved - the boundaries of the objects are known, therefore learning of the CRF is not necessary. Instead we will use the computed features directly. For the same reason the pairwise feature $S(a_x, a_{x'}, I)$ is not used, as its purpose was to ensure emergence of areas labelled in the same way.

The computation if center-surround histogram is modified, such that the $f_h(\cdot,I)$ map is computed on a per object basis, with the rectangle R being smallest bounding box covering an investigated object. The surrounding rectangle R_s is computed as simple extension of the rectangle R. However, wherever possible, gold standard polygons enclosing objects where used instead of bounding boxes to ensure better estimation of histograms.

Finally the features are combined into single energy value as defined by equation 7.6, which is simply assumed to be measure of intrestingness of the considered object.

7.4.3 Color histograms

In addition to the methods described above, we implement and investigate an approach inspired by work on the representation of an object's appearance in computer vision

and spatial color distribution used by Liu et al. (2011). Our model is based on a simplified *Factored Shapes and Appearances* (FSA) representation (Eslami and Williams, 2011). The central assumption of the representation is that the pixels corresponding to each object have been generated by W fixed Gaussians in a feature space. We found *Lab*-space to be the most effective in our initial experiment; see also Dziemianko et al. (2011a).

In first phase the means μ and covariances Σ of these Gaussians are extracted by fitting a Gaussian Mixture Model (GMM) with W components over all pixels in the image. At this stage object boundaries and locations are ignored. In subsequent step, pixels are clustered into W clusters according to the associated GMM components by selecting component $\hat{w}$ that maximizes probability of a pixel being drawn from the Gaussian distribution with mean μ_w and covariance Σ_w:

$$\hat{w} = argmax_w \frac{1}{(2\pi)^{k/2}|\Sigma_w|^{1/2}} e^{-\frac{1}{2}(x-\mu_w)^T\Sigma^{-1}(x-\mu_w)} \tag{7.17}$$

where x is feature vector representing a pixel, while k dimensionality of this vector. The value of W was set experimentally to 15; a similar value was also used by Eslami and Williams (2011).

The final step of the first phase consists of computing global histograms H of the pixel assignments $\hat{w}$. Each histogram is then normalized, dividing each bucket count by the total number of pixels, so that it represents the proportion of pixels belonging to each cluster rather than absolute counts. The whole process is shown in Figure 7.4.

The saliency map is created in the second phase. At this stage the model assumes that the image is fully annotated (i.e., boundaries for each object within the scene are provided). For each of the objects o_i an additional histogram h_i is computed considering only the pixels and their assignments $\hat{w}$ within the boundaries of the object. The histogram h_i is also normalized by the total number of pixels within the object. Histograms computed this way are distributions over the different pixel types present in the scene.

In the following step an *interestingness* value I_i is assigned to each object o_i. In preliminary experiments, we used the Kullback-Leibler (KL) divergence between local (object) pixel distribution h_i and the global distribution H:

$$I_i = D_{KL}(h_i||H) = \sum_{w=1}^{W} h_i(w) \log \frac{h_i(w)}{H(w)} \tag{7.18}$$

The KL divergence measures the expected number of extra bits required to encode samples from h_i when using a code based on H; intuitively, it represents how different

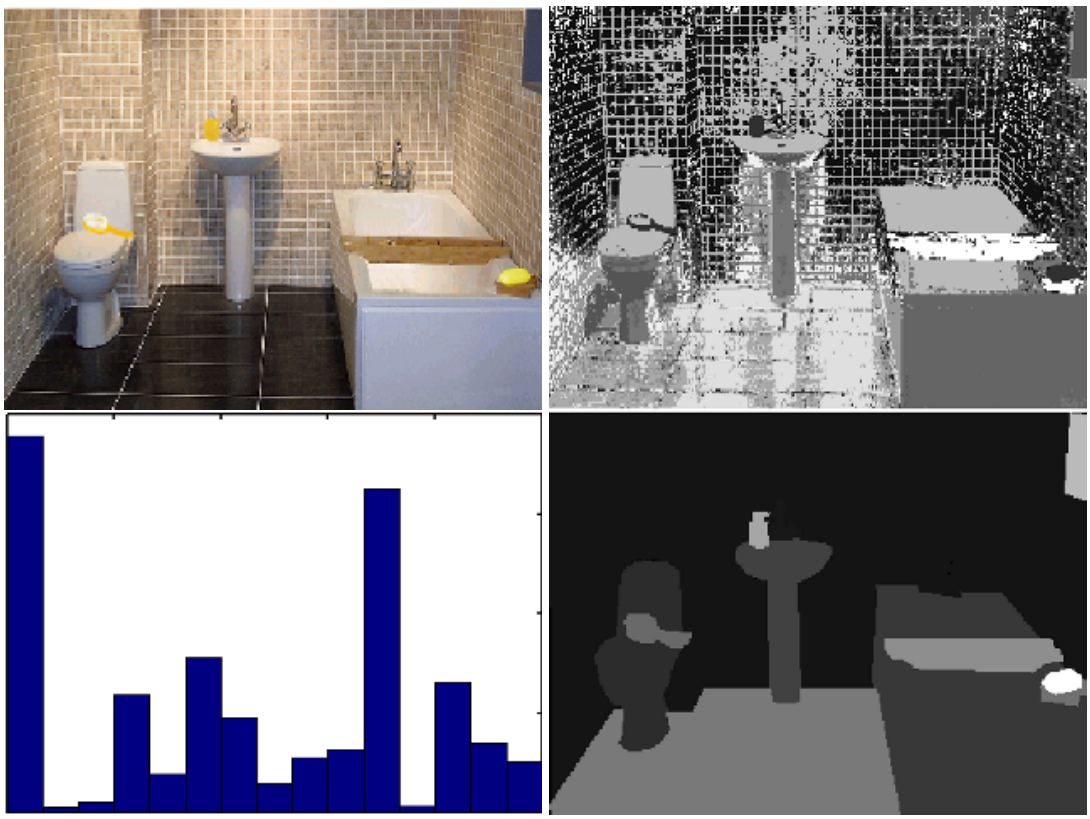


Figure 7.4: Calculation of the global histogram H: from left to right: original image, clustering of pixels to different Gaussian components, histogram of the assignments, and objects interestingness map

the object is from its surroundings (and thus interesting), with a larger value meaning more interesting. The saliency map is constructed by filling the area corresponding to each object with the interestingness value I_i assigned to it.

We also investigate a version of the model, where the interestingness I_i is calculated as a divergence between histogram of local pixel values h_i and histogram of pixel values in object surroundings H_i^s instead of global histogram H:

$$I_i^s = D_{KL}(h_i||H_i^s) = \sum_{w=1}^{W} h_i(w) \log \frac{h_i(w)}{H_i^s(w)} \tag{7.19}$$

The surroundings S^i of the object o_i are computed by expanding the object boundaries in all directions by a specified distance. We found experimentally, that it is beneficial to adjust the distance according to the object's size, thus we perform image dilation operation with rectangle of size equal to objects bounding box:

$$S^i = T^i \oplus R^i = \bigcup_{b \in R^i} T_b^i \tag{7.20}$$

where T^i is a binary mask of the object and R^i is a rectangle of size matching the object bounding box. The dilation operation can be thus seen as locus of the points covered by the rectangle R^i as its centre is moving inside the object.

In experiments described further we also use interestingness value I_i defined as *cross-entropy* between local and global histogram H:

$$I_i^e = -\sum_{w=1}^{W} h_i(w) \log H(w) \tag{7.21}$$

Similarly for surroundings histogram H_i^s:

$$I_i^{e,s} = -\sum_{w=1}^{W} h_i(w) \log H_i^s(w) \tag{7.22}$$

7.5 Object saliency and prediction of eye movements

Traditionally, saliency is used to predict likely fixation locations. Recent studies such as Einhauser et al. (2008); Nuthmann and Henderson (2010) show however that early saliency is not as good predictor as object position. As shown in sections 4.5 and 7.3, using only object position does not result in better predictions, which suggests that, even though the fixations are directed to objects centres, some mechanism is employed to select only certain objects to be attended.

We investigate the hypothesis that object based saliency is a part of such a selection mechanism. We compare it against regular saliency on the Visual Count and Object Naming datasets using the receiver operating characteristic (ROC). Additionally we test two baselines that do not use saliency in any form: The first one weights objects by their euclidean distance from the center of the image, normalized by object area. This approach is inspired by experimental evidence of center bias in scene viewing (Tatler, 2007), and will be referred to as *center bias*.

Secondly, based on the findings of Nuthmann and Henderson (2010), we also include a baseline that predicts fixations by selecting object centers. In this case, a map is built as a sum of Gaussians centered on the bounding boxes of the object in the image. The parameters of the Gaussian are fitted using 10-fold cross-validation to avoid overfitting the datasets. This baseline is referred to as *object overlay.*

The ROC curve is estimated as follows. Firstly the objects in the image are sorted according to their interestingness score. Than the most interesting objects are selected until a desirable area of the image is covered. It is important to note that achieving the exact specified area coverage is not possible, as only entire objects can be selected.

Model	Obj. counting	Obj. naming
Saliency	61.66	55.87
Object overlay	63.60	59.78
Center bias	68.02	69.17
Converted (max)	55.27	64.66
Converted (mean)	70.44	68.65
Liu et al. 2011 features	66.67	67.42
Color-component hist.	66.73	67.40

Table 7.2: Estimated area under ROC curves presented in figure 7.5. All the values are statistically significantly different from regular saliency with $p < 0.01$

The fraction of fixations falling onto the selected area is then calculated and projected onto the plot.

The results are presented in Figure 7.5. The ROC curves show that selection based on object overlay is better than saliency for thresholds smaller than 40%. Object-based saliency models in turn outperform object overlay. Center bias turns out to be a very competitive baseline, which is only matched by converted (mean).

An analysis of the areas under the ROC curves, summarized in Table 7.2, confirm these observations. The ANOVAs reveal that for both datasets, object position overlay is significantly better than saliency with $F(1,24) = 9.27, p < 0.005$ for object counting, and $F(1,23) = 9,84, p < 0.005$ for object naming.

The calculation of area under ROC curve for object-based models is not trivial due to the discontinuity of the plot. We estimated the AUC by interpolating the missing values.[1] The analysis of the interpolated curves shows that for both datasets, object-based selection is superior to traditional saliency, and to object overlay. These differences are statistically significant, for example converted (mean) is better than saliency with $F(1,24) = 165.60, p < 0.001$ for counting and $F(1,23) = 279.30, p < 0.001$ for naming; for color histogram the values are $F(1,24) = 34.67, p < 0.001$ and $F(1,23) = 227.40, p < 0.001$ respectively.

The pattern for Converted (max) is more complicated, however. On the counting data, it is significantly better than saliency ($F(1,24) = 132.10, p < 0.001$), but not

[1] The discontinuities were interpolated by plotting linear segments between end points of the ROC curve.

as good as any of the other methods. On the naming data, it is significantly weaker than standard saliency ($F(1,23) = 245.70, p < 0.001$), operating around chance level. This can be explained by the fact of saliency being sensitive to high contrast edges, usually corresponding to object boundaries. As such, the highest saliency values corresponding to the object might not fall within the object, but rather belong to one of its neighbours.

A surprising results is that object-based selection does not outperform selection based on center bias. However, closer investigation of the object rankings based on center bias and converted (mean) reveals that the average correlation coefficient between the respective rankings is only 0.5 for the naming and 0.43 for the counting data. This indicates that different sets of objects are selected by the two model for a given threshold, accounting for different subset of fixations. A models that combines object-based saliency and center bias there would be a promising next step.

The findings of Nuthmann and Henderson (2010) suggest that it would be beneficial to select the centres of the objects only instead of the whole object in the models described above. Although this is certainly possible, it would impede the main goal of this work: to build object-based framework by converting back to bottom-up, low-level, pixel-based map rather than calculating object interestingness scores directly. We therefore do not investigate this possibility.

In addition to the fixation prediction we investigate correlation of objects ranking with an amount of fixations they receive. We considered only fixated objects to check, whether association between object-based saliency values and total duration of fixations directed towards an object exists. This is different from prediction fixation locations, which can be seen as a binary classification problem (i.e. is object fixated or not), rather than ranking problem (i.e. is it fixated for longer than another object). The comparison of two rankings is done with Spearman correlation coefficient. Table 7.3. summarises the results for the two different data sets.

Surprisingly, the simple maximum value of pixel-based saliency leads to the strongest positive correlation. Also, the center-surround methods lead to significant correlation of the rankings. On the other hand, the mode of pixel-based saliency and divergence between local and global histograms of colors lead to quite strong negative correlations. Again, this is surprising as these were good predictions of whether object is fixated or not (see figure 7.5). These correlations while are not intuitively compatible with the results shown in figure 7.5, can be explained by the difference in the performed task - determining the order in which objects are fixated rather than

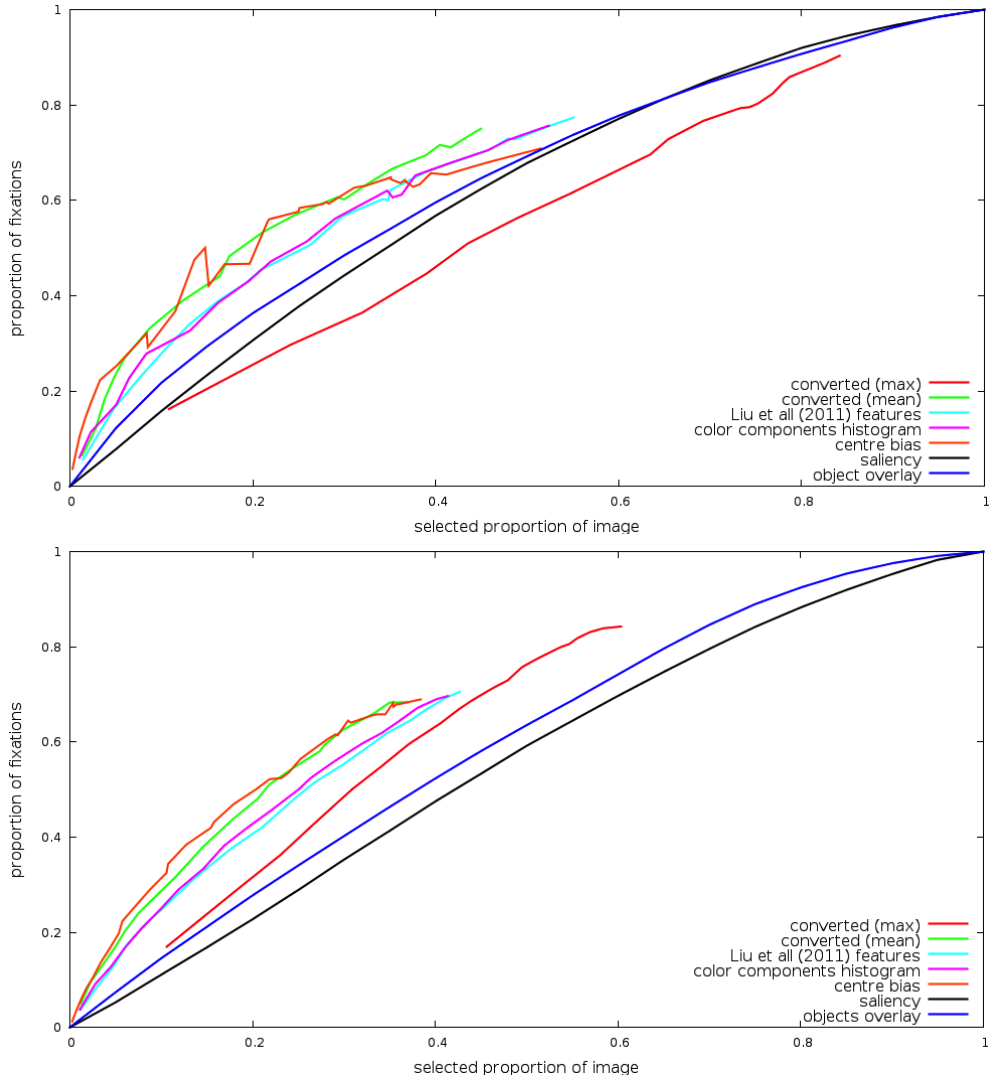


Figure 7.5: Performance of object based selection of fixation locations on the Visual Count (top) and Object Naming (bottom) datasets. Only selected models are shown. It is important to note that traditional saliency and object based models cannot be compared directly due to differences in the selection method. Nonetheless the chart gives a good indication of the relative performance. *Converted* refers to methods described in section 7.4.1, *Liu et al. features* in section 7.4.2, while *colour component histogram* in 7.4.3.

determining whether they are fixated or not. Moreover the procedure applied in the prediction of fixation locations allows for a large number of objects to be selected at the same time. The smaller objects that often have high saliency value (e.g. due to their distinctness from surroundings) do not affect the selected area size as much as larger objects. For example, in the image presented in figure 7.6, 4 out of the top 5 objects

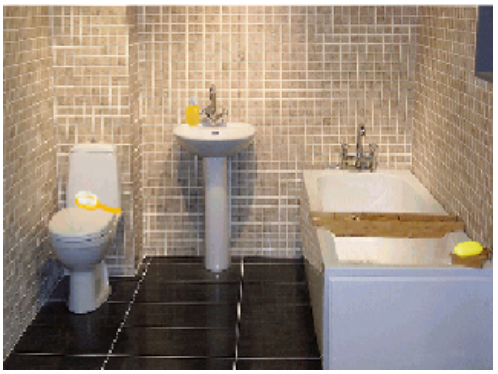
	Object Counting	Object Naming
converted (max)	0.38 (48.61%)	0.36 (54.00%)
converted (mean)	0.01 (20.83%)	-0.06 (16.00%)
converted (median)	0.01 (22.22%)	-0.06 (17.00%)
converted (mode)	-0.14 (26.39%)	-0.21 (35.00%)
histogram KL-divergence N=20	-0.24 (38.89%)	-0.24 (55.00%)
histogram cross-entropy N=20	0.02 (22.22%)	-0.06 (22.00%)
center-surround KL-divergence N=25	0.28 (44.44%)	0.21 (31.00%)
center-surround cross-entropy N=25	0.29 (41.67%)	0.23 (32.00%)

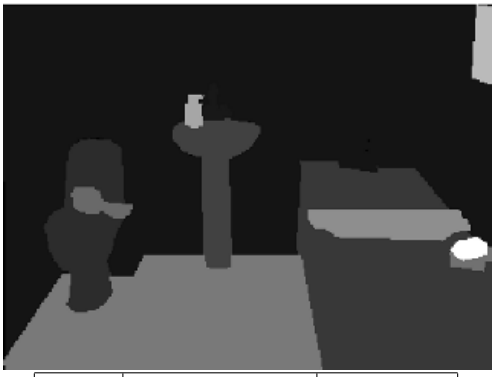
Table 7.3: Average correlation between object rankings based on their visual interestingness score and amount of fixations they receive. The value in parenthesis denote amount of trials where correlation was statistically significant. *Converted* refers to methods described in section 7.4.1, *center-surround* in section 7.4.2, while *histogram* in 7.4.3.

covered less than 1% of scene area. They do, however, contribute the same weight to the rankings. As a result, a large number of small objects at the top of the ranking causes the correlations to drop if they do not have the same ranks in the compared lists. On the other hand, they affect the prediction of fixation locations only for very small threshold values.

In addition to the average correlation score, we report amount of trials on which the correlation was statistically significant with 90% confidence. This is important as the critical value above which the score is significant changes depending on length of compared ranked lists. The length is in our case not constant as it directly reflects number of objects present in the images. Moreover the mean value itself does not provide the whole picture of the values distribution, which might not be uni-modal.

This result shows, that even though certain methods might not be good in discriminating what was fixated and what was not, they can account for other characteristics of human behaviour. It also shows, that modelling of attention requires extensive exploratory study of the data, and robust evaluation procedure to avoid missing important effects.





rank	object	area [%]
1	soap	0.35
2	soap (liquid)	0.29
3	window	0.87
4	container	2.39
5	brush	0.59

Figure 7.6: An example of situation where top saliency ranked objects do not contribute much to area based selection. From top left: an image, object saliency map, selection based on top 5 objects, and areas of individual objects covered by the selection.

7.6 Object saliency in models of synchronous processing

The notion of object saliency as discussed above is not of the main interest of this thesis. Rather, we are rather interested in modelling human eye-movements, ultimately at a scan-path level.

The chapters 5 through 6.4 have discussed the problem of scan-path generation without considering any visual information, while multiple studies mentioned through this work have provided considerable evidence of such information being used by our attentional system. Our own experiment presented in chapter 4 shows cooperation between context and visual features in attentional guidance.

The work presented in previous chapters can be easily extended with the notion of saliency discussed in his chapter. We will start with a discussion of Contextual Guidance Model (see chapter 4) in an object based setting. Than we will follow with an extension of mathematical formulation of models presented in chapter 5.

7.6.1 Contextual guidance in object-based setting

7.6.1.1 Formulation

The Contextual Guidance (CGM) and Memory Modulated Saliency (MMS) models discussed earlier in chapter 4 can be transformed into an object-based framework by simple re-interpretation of their components.

Recall that the overall structure of these models is:

$$M(x,y) = S(x,y)^{\omega} \cdot C(x,y) \tag{7.23}$$

where $M(x,y)$ is a final modulated saliency value for point (x,y), ω is a weight parameter, while $S(x,y)$ and $C(x,y)$ are saliency and contextual bound values at this point.

This framework can be directly applied to objects:

$$M(o_i) = S(o_i)^{\omega} \cdot C(o_i) \tag{7.24}$$

where $M(o_i)$ is modulated saliency score for object o_i, with $S(o_j)$ and $C(o_i)$ being object saliency and contextual scores.

The scores in this framework can be interpreted as likelihoods of the object being fixated based on corresponding set of features. In such a case, we can simply substitute saliency score $S(o_j)$ for saliency score calculated for the object with one of the models discussed earlier. The value $C(o_j)$ can be substituted with the appropriate contextual fitness estimated for the object's position. As we would consider the object as a whole entity, this value would be calculated only once for the objects' centre, rather than as in the original setting for each of the image pixels.

Following CGM of Torralba et al. (2006) we can define $C(o_j)$ as:

$$C(o_j) \quad \propto \quad P(X|O=1,G) \tag{7.25}$$

where X is the objects position, G is set of image global features, and O is a binary variable denoting presence of the object in the scene[2].

The other details of the model - namely training and inference of probability $P(X|O=1,G)$ do not need any major changes and can be used directly as they are described in Torralba et al. (2006).

It is possible to modify the MMS model following the same concept - substitution of pixel-based saliency and contextual values with their object-based equivalents.

[2]It is important to remember, that the CGM model was developed to model search task in naturalistic scenes, where the target object might not necessarily be present. The probability $P(X|O=1,G)$ should be therefore interpreted as a probability of presence of a target object at a given position X

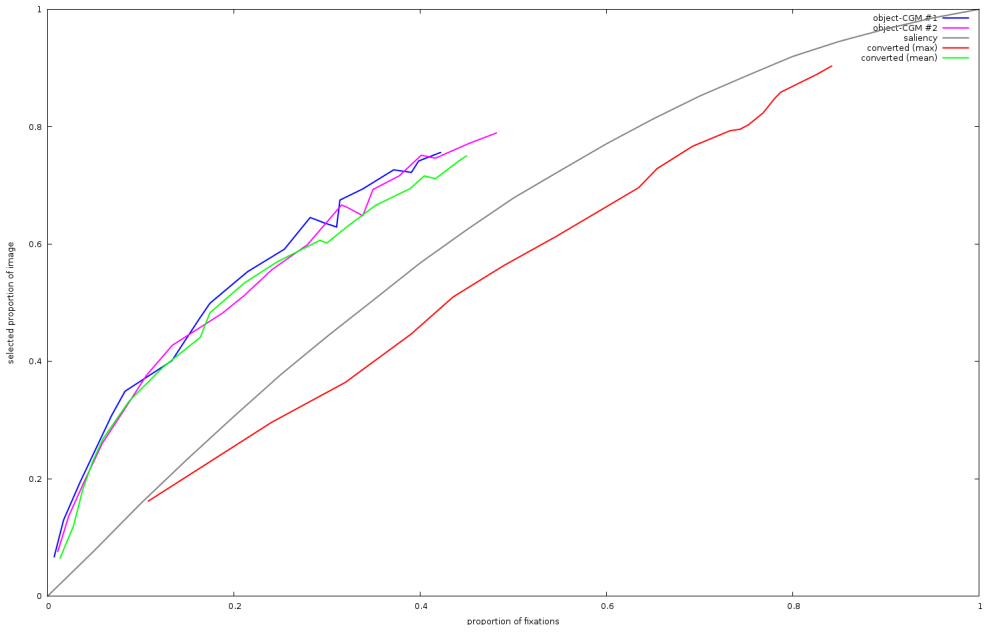


Figure 7.7: Evaluation of object-based Contextual Guidance Model on Visual Count data. For clarity presented are only some representative samples: regular saliency, saliency converted into object based representation and two CGM models using this saliency with $\omega = 0.5$ and 0.25.

7.6.1.2 Evaluation

We evaluate the object-based CGM models on the Visual Count datasets used in chapter 4. We do not use the original dataset of Torralba et al. (2006) due to lack of appropriate object annotations.

The results are summarised in figure 7.7. The presented results reveal a trend similar to the one achieved with traditional, pixel-based model. A small improvement over the saliency in terms of AUC is achieved: 72.10% and 71.70% for object-based CGM with $\omega = 0.5$ and 0.25 respectively compared to 70.44% for object saliency alone. These results are however not statistically different ($p > 0.05$)

We can theorize, that improvements on the original dataset would be considerably higher given simpler task, search rather than count, and less variability in the stimuli, like it is a case for regular CGM and MMS models (see chapter 4).

7.6.2 Scan-paths prediction

7.6.2.1 N-gram based scan-path generation

The models capable of generating human-like scan-paths presented in chapter 5, as well as the shallow parsers from chapter 6 can easily be extended to use information about saliency.

In further discussion we will define saliency as likelihood (or proxy of thereof) of an object being fixated based on its visual features. This assumption is a key factor simplifying the analysis. Even though the calculated scores are not necessarily probabilities, they can easily be converted into them, either by appropriate renormalizing, or by building simple probabilistic model $P(f_o = 1|S(o))$ where f_o is a binary flag indicating fixation at object o, while $S(o)$ is saliency score of this object.

The model described by equation 5.6 in its current form, assumes the probability of an object being fixated being conditioned on previous fixation and current syntactic interpretation of a sentence. This can be easily modified by conditioning on some additional factors - such as discussed saliency.

Formally we can define the model by rewriting equation 5.6 as follows:

$$\begin{aligned} P(O|S) &= P(o_1...o_N|s_1...s_M, S(o_1)...s(o_N)) && (7.26)\\ &\propto P(len(O) = N) \cdot \\ &\quad \prod_{i=1}^{M} P(len(O_{I(s_1...s_i)}) = n_i) \cdot \\ &\quad \prod_{i=1}^{M} \prod_{o_j \in O_{I(s_1...s_i)}} P(o_j) \cdot P(o_j|o_{j-1}, I(s_1...s_i), S(o_j)) && (7.27) \end{aligned}$$

where $O = (o_1, ..., o_N)$ is a sequence of fixated objects (scanpath), $S = (s_1, ..., s_M)$ an encoded sentence, o_i and s_i i-th object and sentence part respectively, O_{s_i} is a subsequence of O at sentence part s_i, while N and M total length of scanpath and sentence.

Based on our assumption of saliency being proxy for likelihood of the object being fixated conditioned on its visual features, we can further write:

$$P(o_j|o_{j-1}, S(o_j)) = \frac{P(S(o_j), o_j, o_{j-1})}{P(S(o_j), o_{j-1})} \quad (7.28)$$

which can be estimated from the data.

Alternatively we can modify the prior distribution $P(o_j)$, such that $P(o_j)^* \propto$

$P(o_j|S(o_j))$, and:

$$\begin{aligned} P(O|S) \quad \propto \quad & P(len(O)=N)\cdot \\ & \prod_{i=1}^{M} P(len(O_{I(s_1\ldots s_i)})=n_i)\cdot \\ & \prod_{i=1}^{M}\prod_{o_j\in O_{I(s_1\ldots s_i)}} P(o_j)^*\cdot P(o_j|o_{j-1},I(s_1\ldots s_i)) \end{aligned} \tag{7.29}$$

7.6.2.2 Scan-path shallow parsers

Similar approach can be used to modify syntactic chunkers from chapter 6. In this case the probability of a parse is tied with a particular object j through emission probability $b_{c(t)}(o_j)$ where $c(t)$ is HMM state at ime t, and o_j denotes a fixation a object j. Intuitively this probability would be dependant on the saliency of the object j, hence:

$$b_{c(t)}(o_j) \quad \propto \quad P(o_j|S(o_j)) \tag{7.30}$$

which can again be incorporated into learning algorithm and estimated from the data. Another, much simpler approach can be applied to Two-Output and Coupled HMM. As these models have been designed to handle multiple observable streams, including prosodic annotations (see Pate and Goldwater, 2011), it would be more natural to apply the same approach. In such a setting, visual saliency is modelled as an additional stream of values accompanying the scan-path and sentence. The parsers are therefore extended to handle three, rather than two, streams of observable variables as depicted in figure 7.8. The joint probabilities of such models are given by following equations:

$$P(C,S,F,O) \quad = \quad a_{c(0)c(1)}\prod_{t=1}^{T-1} b_{c(t)}(f_t)d_{c(t)}(s_t)e_{c(t)}(o_t)a_{c(t)c(t+1)} \tag{7.31}$$

being equivalent of model from equation 6.2, where again $C = c(0),...,c(T)$ is the sequence of states (effectively chunk tags), $a_{c(t)c(t+1)}$, $b_{c(t)}(f_t), d_{c(t)}(s_t), e_{c(t)}(o_t)$ transition and emission probabilities, while $S = s_1,...,s_T$, $F = f_1,...,f_T$ and $O = o_1,...,o_T$ are the sentence, scan-path, and sequence of fixated object features.

Similarly we can develop equivalents of the Coupled HMM from equation 6.3, with multiple possible internal architectures.

The training of such models is conceptually no different than for two-output case.

Although these modifications seem straightforward, one complicating factor exists - the lack of a suitable dataset to develop and evaluate the models. The comprehension

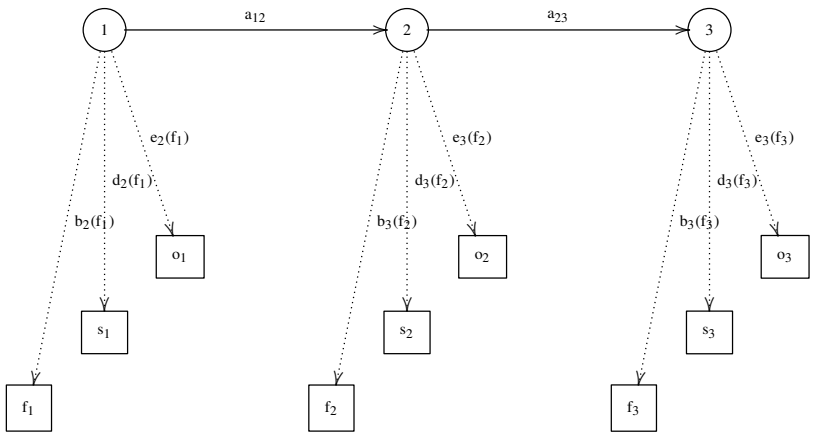


Figure 7.8: An graphical model of architecture of shallow parser that uses additional information about object appearances to perform syntactic chunking of scanpaths.

dataset described in section 3.3 and used through this thesis is unsuitable for two reasons. Firstly, the stimuli consists of visual arrays - and as a result the saliency of any object present in the images is very high due to its distinctiveness from the background. Even though modifications of saturation and luminance have been applied to some of the objects, they are still significantly different than the background. Small number of objects present in the images further emphasize his problem.

Secondly the dataset is too small to reliably estimate extended list of model parameters.

As a result we can not perform appropriate evaluation - other available datasets are either affected by the same set of problems, or introduce new ones such as very weak referential correspondence between linguistic and visual stimuli.

7.7 Saliency and prediction of high-level perception responses

The results presented so far show that object-level visual features can partially explain the selection of objects to be fixated. At the same, time the results of various experiments show a relatively high level of consistency between subjects. This includes not only which objects fixated, but also the responses to experimental tasks such as object naming.

In this section we aim to investigate whether saliency has an effect on our behaviour in variety of tasks. However, we do not assume saliency is correlated with

fixation locations - we consider a possibility, that salient areas might not be fixated at all. This can happen due to multiple reasons. Firstly, some other factors such as context, memory, and task constrains might have stronger effect on our behaviour, limiting the influence of saliency. Secondly our vision system is quite robust and allows us to access and process a fair amount of information without directly fixating some image areas by means of mechanisms such as para-foveal vision, gisting, or by using non-visual information. For example Biederman (1972); Potter (1975) and others have shown, that humans are able to understand visual scenes quickly and easily, often with only one fixation. Moreover color and texture (Oliva and Schyns, 2000), objects (Biederman et al., 1982; Wolfe, 1998), and scene layout (Biederman et al., 1974; Schyns and Oliva, 1994) are extracted or inferred during initial scene inspection before high level visual processing. Potter (1975); Potter et al. (2004) show, that scenes matching high level, conceptual description (such as "image of a birthday party") are easily and consistently detected by human observers in a stream of rapid displays, even if each scene in the sequence is shown for less than 100ms.

We theorise, that even though saliency does not necessarily strongly correlate with fixation locations, it might have an effect on other aspects of our behaviour. This is not an entirely new hypothesis - for example Spain and Perona (2011) present a formulation of object importance, that takes into account multiple visual features. Similarly Liu et al. (2011) presents a system capable of detecting objects perceived as salient by humans in uncluttered images.

We also hypothesize that saliency should be treated as a feature of an object, rather than a location or area. This follows results of studies such as Nuthmann and Henderson (2010) presenting strong evidence for object based allocation of attention. Similar assumptions have already been made in past work on attentional systems in robotics (Yu et al., 2010; Schauerte et al., 2012). However, these implementations rely on so-called *proto-objects* (see e.g. Walther and Koch, 2006) that should be seen as related more to segments computed with algorithms such as *normalized cuts* (Shi and Malik, 2000), rather than true objects, and were shown not to be basis for allocation of attention by Nuthmann and Henderson (2010) as already discussed earlier.

7.7.1 Experiment

To investigate, whether object saliency as discussed above has an influence on high level perception, we used the datasets described in chapters 3.4.1 and 3.4.2. The results

	salient	interesting	important	naming results
eye-catching	0.22 (18.56%)	0.24 (16.49%)	0.01 (18.56%)	0.04 (6.00%)
salient		0.15 (8.16%)	0.04 (14.29%)	0.04 (4.00%)
interesting			0.03 (13.27%)	0.06 (5.00%)
important				0.03 (5.00%)

Table 7.4: Average correlations between the object rankings in each of the critical conditions and results of object naming. The number in parentheses indicated percentage of images where achieved correlation was statistical significant.

of the experiments (i.e. rankings of objects for various questions) were fed into an urn model (Spain and Perona, 2011), that allows us to rank the objects with respect to the probability at which they are chosen by participants at a given trial. The model can be seen as a process of drawing a ball from an urn without replacement. The urn contains one ball for each object type in the image, and its size affect probability of being chosen. The draws are assumed to be independent - essentially disconnecting the mentioned objects from each other.

The urn model also assumes the same starting condition for all samplings - that can be intuitively seen as lack of use of behavioural data (eye-movements) to shape probability distributions. Moreover the balls are removed from the urn only if they are drawn. These two assumptions are a little limiting as we can reasonably expect the importance of the objects to be dependent on the interpretation of a scene, and that some object are not going to be mentioned in conjunction with other (e.g. it seems intuitively unlikely for 'face' to be mentioned after whole 'head' being named).

The model parameters are estimated as the Maximum-Likelihood fit to training experimental data with respect to the probability of observing a set of mentioned object sequences. Details of the parameter estimation and its derivation is given in Spain and Perona (2011).

7.7.1.1 Results

We start the analysis of the results by comparing the scores assigned to the objects for each critical question. Table 7.4 presents correlations between the objects selected by participants of Mechanical Turk experiment as they are grounded to gold standard annotations.

Not surprisingly, the rankings obtained for most 'eye-catching', 'salient' and 'interesting' are to some extend correlated with each other. The weaker correlation of 'interesting' with the others categories might be explained by subjects taking into account other factors than visual features, such as contextual fitness or familiarity, while determining the ranking.

The responds for *important* are on average not correlated with the responses to questions focusing on visual features. This is in contrast to the model of Spain and Perona (2011), which extensively uses visual features to determine object importance. This is not surprising taking into account cognitive relevance framework. It is however important to note that importance were determined by the subjects without any particular task in mind, allowing broad interpretation of what was most important object.

Quite surprisingly, the responses to the Mechanical Turk tasks are not similar to the results of Object Naming task with the correlation coefficients close to zero. Our initial suspicion was, that this result might be related to the fact that the object naming responses do not necessarily have a one-to-one correspondence with objects in the image. We have ruled out this possibility by extracting all polygons that are either representing object enumerated by the subjects, or object that is semantically related (e.g. saucer while cup was mentioned), and calculating fraction of named objects, for which at least one polygon is on the list of objects selected in investigated Mechanical Turk task. These results can be found in table 7.5. It is visible that only a relatively small amount of objects can be found on both lists, regardless of investigating all possible one-to-many relations between named objects and polygons representing objects in the images. This suggests, that the naming task is not guided by saliency or importance as it is perceived at high cognitive level. Possibly the low overlap between the results is a result of task differences - object naming seems to be more exploratory than mechanical turk tasks.

Table 7.6 presents correlations between the responses to the questions with various visual features. It can be seen that the maximum saliency value within an object's boundary is the strongest correlated feature with responses to all the questions, even though the amount of trials where the correlation was significant is much lower for *interesting* and *important* than for *salient* and *eye-catching*.

In general the correlation between visual features and responses are similar to the correlations with eye-movements presented in table 7.3. The only exception is the correlation between local and global histograms KL-divergence and responses to *eye-catching* and *salient*. Additionally, some negative correlations are stronger in case

Question	Fraction [%]
eye-catching	12.89
salient	11.44
interesting	13.01
important	12.28

Table 7.5: The fraction of objects mentioned in naming experiment, that might be grounded to objects selected in Mechanical Turk experiments

	eye-catching	salient	interesting	important
converted (max)	0.23 (19.61%)	0.18 (21.00%)	0.21 (13.00%)	0.23 (12.00%)
converted (mean)	-0.01 (14.00%)	-0.03 (16.00%)	0.04 (9.00%)	-0.09 (15.00%)
converted (median)	-0.04 (14.00%)	-0.04 (16.00%)	0.06 (9.00%)	-0.10 (14.00%)
converted (mode)	-0.11 (17.00%)	-0.10 (18.00%)	-0.11 (15.00%)	-0.19 (18.00%)
histogram KL-divergence N=20	0.00 (15.00%)	-0.11 (12.00%)	-0.06 (10.00%)	-0.20 (12.00%)
histogram cross-entropy N=20	0.07 (17.22%)	-0.07 (13.00%)	0.01 (11.00%)	-0.04 (14.00%)
center-surround KL-divergence N=20	0.18 (21.00%)	0.09 (14.00%)	0.11 (10.00%)	0.14 (11.00%)
center-surround cross-entropy N=20	0.19 (13.00%)	0.18 (16.00%)	0.17 (15.00%)	0.11 (17.00%)

Table 7.6: Average correlation between object rankings based on their visual interestingness score and responses to the experimental questions. The value in parenthesis denote amount of trials where correlation was statistically significant.

of responses to *important* than to other cases suggesting that the objects with some distinct visual features are not considered important. Responses to *eye-catching* and *salient* were correlated with visual features at significant level at larger percentage of trials than those to *interesting* and *important*. It suggest that investigated visual features contribute to the selection of objects at various level, dependant on the task, scene context, and relations between objects.

Calculating a correlation of responses to naming experiment with visual features is not feasible as there is no one-to-one mapping between objects mentioned and objects in the scene. For example 'chair' might occur in scene several times, while being mentioned only once. Similarly objects not present in the scenes are being mentioned due to contextual fitness (e.g. mentioning saucer while only cup is present in the image). Inability to find onefold one-to-one mapping results in multiple possible rankings and as a result multiple possible correlations.

7.8 Summary

In this chapter we have investigated relations between object, its appearance, and allocation of attention. We have demonstrated, that it is possible to apply object-based version of saliency to prediction of fixation locations. Object-based saliency is not calculated as a value for each of the image pixels, but rather over an area within the boundaries of an object. In this approach, saliency is treated as a feature of an object, similar to other features such as position. This approach is compatible with theories assuming an object-based allocation of attention, such as the cognitive relevance framework (Henderson et al., 2007, 2009). At the same time, off-object fixations can be explained by inaccuracies of the oculomotor system (see. e.g. Nuthmann and Henderson, 2010), averaging of population across scene or its part (see e.g. Zelinsky, 2012) or errors of fixation detection algorithms of eye-tracking software as discussed earlier.

Even though the intuition that salience is a property of objects has been utilized before, we are not aware of any previous studies that investigate if object-based saliency can reliably predict human fixations. We showed that the prediction of fixations based on objects and their visual features is not only possible, but superior to standard saliency. We have also found, that popular technique of using the maximum value of saliency within an object was not confirmed as a reliable predictor of whether object is going to be fixated, which is a important result considering the popularity of this feature in previous modelling studies. On the other hand maximum of saliency is a relatively good predictor of the amount of fixations on objects that were actually fixated.

We also discussed a way of incorporating the visual cues into models presented in chapters 5 to 6, opening a way to build a complete model capable of predicting scanpaths in situated language comprehension. However, we were not able to evaluate the developed idea due to lack of appropriate datasets as discussed in section 7.6.2.

Finally we have analysed the experimental data in order to understand how an object's saliency influences higher level cognitive processes. We have found, that objects that are perceived as visually attractive are not necessarily perceived as important part of the scene. Moreover neither of these correlate strongly with results of object naming. At the same time all of them are weakly correlated with some of the object-saliency models discussed. It it a set of interesting findings on one hand confirming, importance of the contextual task factors, and on the other the influence of appearance on the attentional selection.

Chapter 8

Summary and Future Work

In this chapter we summarise the main contributions of this thesis.

Our primary goal through this work was to investigate the possibility to predicting and generating human eye-movements. A particular focus was put on scan-paths as a representation whose importance is often underestimated.

Furthermore we studied the *object* as a target of attentional selection along with related issue of whether this selection is governed by bottom-up or top-down process.

In this thesis we do not present one complete and coherent cognitive theory, but rather we build a set of models fitting modern frameworks, like e.g. Cognitive Relevance Theory (Henderson et al., 2007), allowing for the study of scan-paths and synchronous multi-modal processing in general by providing an explanation of several phenomena observable in experimental data.

8.1 Contributions and their implications

In the first part of this thesis we re-visited a classical model of top-down contextual guidance in visual search of Torralba et al. (Contextual Guidance Model (CGM) 2006). We have hypothesized that a similar effect might be achieved through on-line accumulation of recent experience. As a result we presented a computational model - Memory Modulated Saliency (MMS) - that improves over bottom-up saliency showing the benefit of memory for fixation prediction, without necessity for calculation of any image or scene statistics. Instead it relies on memory of previously seen target positions.

It also achieves performance comparable with the CGM, however does not require any off-line training data in the form of annotated images, but incrementally learns likely target positions. This means, the model can easily adapt to new datasets, tasks,

and experimental conditions - unlike CGM which is sensitive to the nature of the training data.

Moreover we have shown, that the predictions of the two models are not equivalent, and combining them using a weighted sum is beneficial, resulting in further improvement of fixation location prediction. This result suggests, that a complete model of attentional guidance needs to combine features of both models, that is contextual guidance based on scene properties, and memory of recent experience.

The MMS model presents an approach that is conceptually different than CGM. The CGM, by assuming an offline learning phase based on large training set, effectively models how humans learn the typical positions and associations of objects during the lifespan. The MMS models short-term learning as it happens while a human performs a specific task. It seems that human behaviour is driven by both types of learning, and the two models should be seen as complimentary, emphasizing the need for a model that integrates components from both the CGM and the MMS.

Subsequently, we investigated if the saliency modulation framework used in construction of CGM and MMS models can be used to predict fixation locations in tasks with incrementally evolving stimuli - particularly Visual World Paradigm speech comprehension experiments. We presented a model, which performs modulation of saliency with an overlay map dependent on interpretation of a sentence at a given time. The overlay map is computed as a sum of Gaussians centred at the objects' locations weighted by probabilities of the objects being fixated during the time period the map is computed for.

The resulting model is better than saliency in selection of objects that are likely to be fixated at the investigated time period. The probabilities used to compute overlay and applied to saliency were learnt from experimental data, which in conjunction with the results suggests, that scan-paths carry interesting temporal information, that might be extracted and used to predict attentional shifts during sentence interpretation.

In addition the results indicate, that an interpretation of a sentence at a given time is crucial to build successful models of the eye-movements, while considering the sentence one word at a time is not sufficient. This is especially true in the presence of referential ambiguity, as it is the case in experimental materials used in evaluation of the models.

This findings confirm the second of the claims of this thesis, of the scan-paths being influenced by the interpretation of the linguistic input, and indicate strong need for attentional models to include powerful contextual components.

The second part of the thesis investigates the scan-paths in more depth. Based on results mentioned above, we hypothesized, that they are highly structured, and carry interesting information that might be extracted and used in the construction of models of synchronous multi-modal processing.

The result of the work is a series of models capable of generating human-like sequences of fixated objects given the Visual World Paradigm stimuli consisting of a sentence and a simple representation of a visual scene. In contrast to existing studies, we aimed to model full human-like scanpaths, rather than solve tasks such as anticipation of linguistic material (as e.g. Mayberry et al., 2005). Moreover we applied the models to more complex stimuli with referential ambiguity arising from the syntactic and visual material.

We show, that scanpaths are consistent across the subjects to a degree that allows their analysis and synthesis. The results also confirm, that linguistic material and its interpretation at a given time affects the attention and scan-patterns. This confirms our first claim of the the scan-paths carrying an interesting information that can be recovered, modelled, and used.

Based on intuition of scan-paths being structured at a level higher than local proportion of looks, we applied an idea of *shallow parsing* (Abney, 1992) to the sequences of fixations. Using machine learning techniques it was not only possible to construct parsers explaining training data, but what is more important, the parsers were able to generalize well to unseen data.

We have successfully applied the parsers to scan-path synthesis and other tasks such as subject identification. These results confirm, that the structure in scan-paths during situated language comprehension goes beyond simple statistical measures. Moreover it can be extracted, and used in synthesis resulting in sequences of fixations, that are not only within range of differences between humans in terms of Scan-Match distance, but also exhibit similar patterns of behaviour at a global scale.

The parsers were also capable of performing tasks such as subject classification. Despite low number of training samples sourced at different experimental conditions, the performance was superior to the widely used Scan-Match metric. Following this results, we investigate and implement a proof-of-concept method to calculate scan-path similarity. Even in its current, very simple, form the method achieves performance comparable with Scan-Match. The conceptual difference i.e. comparing high level structure rather than calculating optimal alignment leaves space for further work with the possibility of developing a successful similarity metric.

Results obtained with the shallow parsers are further confirming that scan-paths carry consistent information that can be utilised. Moreover they show, that even though human behaviour is to a great degree consistent, and depends on characteristics of the stimuli contains strong component specific to a particular individual. The implications of these findings are very important. Firstly they have to be taken into account while analysing experimental data. The individual differences might shadow common behaviour if the influence of linguistic processing on visual attention is weak and sample is inadequate in size. Similarly an individual behaviour might lead to incorrect conclusions about correlations of speech and eye movements, which will not generalize to other subjects. For example calculating similarity metrics using algorithms such as Scan Match might result in unexpected similarities being found due to patterns emerging from individual behaviour.

Secondly the fact, that scan-path structure has an interesting structure results in possibility, and necessity, to perform analysis not only at coarse, collective level, but also at a level of individual scan-paths. The ability to find whether sequences of fixations match the expected behaviour is a powerful tool, revealing cross modal processing effects when statistical analysis of proportion of fixations, or alignment-based similarity measures indicate only weak, insignificant correlations.

Finally, the dependency of behaviour on interpretation of the linguistic input can lead to interesting, practical applications of the discussed models. For example in aided learning of foreign languages the differences between expected and observed behaviour might be used to assess the level of understanding of sentences. Moreover the observed differences can serve as a guidance to understanding what mistakes the subjects do while interpreting the speech. Similar modelling-based approach to explain errors of school children learning multicolumn addition was presented by Young and O'Shea (1981), proving that such method can be an effective way of studying the cognitive process resulting in erroneous behaviour.

The predicted behaviour can be also used to refine the audio-visual stimuli in order to achieve the desired effect. For example, the speech stream can be modified or timed such that, the visual attention accesses most relevant parts of presented scene, allowing involuntary comprehension of important information. Such a possibility is desirable in the preparation of presentations and lectures to ensure that the viewers do not miss the relevant facts or presented knowledge, and in advertising, where conversion rate might be improved by emphasizing certain pieces of information.

The main drawback of the models mentioned above, is that they do not perform

any visual processing. This problem was addressed in the final part of the thesis by investigation of the *object* as a target of attentional selection. We hypothesized, that saliency should be perceived as a higher-level feature of an object, rather than bottom-up features associated with pixels or areas.

We have shown, that attentional selection based on object saliency can be modelled, and is superior to selection based on classical saliency models like those of Itti et al. (1998). In this setting the saliency is calculated as a feature of an object, rather than as a value at each pixel of an image. This is important evidence in favour of theories of top-down attentional selection. It however does not diminish the importance of visual features, confirming our third claim of the importance of the visual appearance in synchronous processing.

Moreover this interpretation allows simple integration of visual processing path into the scan-path generation process. Specification of such models is discussed, however they were not evaluated due to lack of appropriate experimental dataset. In addition we have shown that visual features might be easily integrated into shallow parsers, resulting in a setup similar to prosodic bootstrapping in Natural Language Processing (see Pate and Goldwater, 2011).

We have also presented applications of object saliency to the classical models such as Contextual Guidance Model (CGM Torralba et al., 2006). We have shown that it is possible to translate these models into object based framework. This is important in order to integrate many classical experimental and modelling studies, that explain certain phenomena crucial to understanding visual attention in depth.

Overall, this thesis presents a complete set of models and tools that allow synthesis of human-like fixation sequences in situated language comprehension. At more theoretical level we show, that scan-paths carry information, which is worth further investigation, and which certainly should not be neglected in future studies of multimodal language processing.

8.2 Future work

Regardless of the contributions of this thesis, a large number of open questions and possible directions of research remains.

The first, most important and immediate is the evaluation of the scan-path generation models incorporating visual features. This is not a trivial issue, as it requires an experimental work and extensive data collection. This work would require a stimuli

that is much more complex than any of the datasets used in this thesis. Particularly, the stimuli should consist of naturalistic, or photo-realistic, scenes, preferably with controlled saliency of the objects. This can be achieved by recreating Coco (2011) experiments with a proper set of images instead of visual arrays.

The second problem to address is extension of the models to not only predict the sequence of fixated objects, but also a duration of each fixation. A considerable experimental evidence of fixation duration being influenced by various factors exists, especially in reading (see e.g. Rayner, 1998, for a review), and are widely acknowledged as an indicator of cognitive load. For example Henderson et al. (1999); Vo and Henderson (2010) discuss the influence of task on fixation durations - which are typically shorter during visual search than memorization. Loftus (1985); Loftus et al. (1992) and others show, that fixation durations are sensitive to degradation of the visual stimuli such as changes in luminance, high-pass filtering etc. Fixation durations are also influenced by contextual and semantic information with longer looks onwards objects that are less consistent with rest of the scene (see e.g. De Graef et al., 1990; Henderson et al., 1999; Hollingworth et al., 2001; Vo and Henderson, 2009).

However, the durations are rarely modelled explicitly in studies of visual attention (see e.g. Nuthmann and Henderson, 2010), even though they are often present in models of eye-movements in reading (e.g. Engbert et al., 2005; Reichle et al., 2003). Incorporation of fixation duration prediction, or more general saccade programming into the models, would result in a more complete and powerful framework.

Similarly the models could be extended, such that, they not only predict which objects are fixated, but which part of an object is fixated - going as far as prediction of actual fixation coordinates. This is not an issue of great importance, as the eye-movements are subject to variability resulting from imperfections of eye-gaze muscular system that cannot be accounted for at cognitive modelling level. However such modelling would require deeper investigation into cognitive processes governing selection of region of interests.

Another related question is whether a speech signal can trigger attentional shift. An audio signal was shown on multiple occasions to be correlated with movements of various body parts - especially face and head. For example head motion of a speaker is related to fundamental frequency (F0) and root mean square (RMS) of the amplitude of speech (Honda, 2000; Yehia et al., 2002), eye-brow movement is related to F0, pauses, and changes to the speech flow (Ekman, 1979; Cave et al., 1996). Moreover characteristics of eye motion change according to the mode of whether it is talking or

listening (Lee et al., 2002). Eye blinking takes place on accented words and pauses (e.g. Condon and Osgton, 1971; Ekman, 1979).

These studies suggest that a wide range of involuntary movements are correlated with speech, and it is reasonable to hypothesise, that certain changes in the speech stream (e.g. of fundamental frequency) might initiate saccadic movements, especially during speech production.

Speech production mentioned above is also an interesting area, that was not investigated in this thesis. We believe, that presented models and tools might also be applied in this case. Some supporting experimental evidence comes from Coco (2011), who shows that characteristics of produced sentences and eye-movements are, to a certain degree, correlated. Such study would however require collection of experimental data in highly controlled environment.

The next area of interest for further research is applicability of the shallow parsers. Their investigation in this thesis was strictly limited to one dataset, therefore the most important issue is to confirm if described findings generalize to other datasets and situations, including tasks other than language comprehension.

The applicability of the parsers to scan-path comparison was briefly studied in section 6.4 with some promising results. As we believe that the problem of comparing scan-paths in a way that takes into account not only local but also global similarity, and is not susceptible to larger but infrequent differences (e.g. occasional re-fixations with long saccadic movement) is not solved, this work should be continued with a focus on developing a technique that allows calculation of distance or similarity between two scan-paths without use of any additional training data.

The initial and final parts of the thesis discuss an issue of top-down attentional selection. Even though the discussion of the object based allocation of attention is - to our best knowledge - the only such attempt thus far, it is rather brief and requires extensive follow-up including appropriate experimental and modelling studies.

The first task is to investigate the predictive power of object based saliency on larger numbers of datasets, preferably collected during various experimental tasks and conditions. At the same time more extensive studies of the possibility of translating and adapting existing classical models into object based framework should be performed.

Moreover the presented evaluation considered only one object feature at a time, while traditionally models of saliency combine multiple such features such as intensity, color and orientation. The logical step is therefore to investigate combination of features presented in chapter 7.

Finally we believe that additional investigation is required to explain the surprising result of object naming and Mechanical Turk experiments not being correlated. This requires replication of the Mechanical Turk experiment in a controlled environment, possibly with simultaneous eye-tracking.

A certain level of further attention might also be given to the Memory Modulated Saliency model discussed in the first part of this thesis. The memory of recent experience in this model is accumulated directly, without access to any image statistic or representation (e.g. gist). The natural extension of the model, would be to allow the memory to not only learn the likely target positions, but also their association with scene layout in a manner similar to this utilised by Contextual Guidance Model of Torralba et al. (2006).

Bibliography

Abney, S. (1992). *Parsing By Chunks*. Kluwer Academic Publishers.

Altmann, G. and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Altmann, G. and Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, 57:502–518.

Altmann, G. and Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, 33:583–609.

Altmann, G. and Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, 30(3):191–238.

Anderson, J. R., Matessa, M., and Douglass, S. (1995). The act-r theory and visual attention. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*.

Bar, M. (2004). Visual objects in context. *Nature Neuroscience Review*, 5:617–629.

Biederman, I. (1972). Perceiving real-world scenes. *Science*, 177:77–80.

Biederman, I., Mezzanotte, R., and Rabinowitz, J. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14:143–177.

Biederman, I., Rabinowitz, J. C., Glass, A., and Stacy, E. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, 103:597–600.

Binder, K., Duffy, S., and Rayner, K. (2001). The effects of thematic fit and discourse context on syntactic ambiguity resolution. *Journal of Memory and Language*, 44(2):297–324.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blake, A., Rother, C., Brown, M., Perez, P., and Torr, P. (2004). Interactive image segmentation using an adaptive GMMRF model. In *Proc. European Conference on Computer Vision*.

Boykov, Y. and Jolly, M. (2001). Interactive graph cuts for optimal boundary and region segemntation of objects in N-D images. In *Proc. IEEE International Conference on Computer Vision*.

Brockmole, J. and Henderson, J. (2006a). Using real-world scenes as contextual cues for search. *Visual Cognition*, 13:99–108.

Brockmole, J. R., Castelhano, M. S., and Henderson, J. M. (2006). Contextual cueing in naturalistic scenes: Global and local contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32:699–706.

Brockmole, J. R. and Henderson, J. M. (2006b). Recognition and attention guidance during contextual cueing in real-world scenes: Evidence from eye movements. *Journal of Experimental Psychology*, 59:1177–1187.

Bruce, N. and Tsotsos, J. (2006). Saliency based on information maximization. In *Advances in Neural Information Processing Systems 18*, pages 155–162. Cambridge, MA: MIT Press.

Castelhano, M. and Henderson, J. (2007). Initial scene representation facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 33(4):753–763.

Castelhano, M., Mack, M., and Henderson, J. (2009). Viewing task influences eye movements during active scene perception. *Journal of Vision*, 9(3):1–15.

Cave, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., and Espesser, R. (1996). About the relationship between eyebrow movements and F0 variations. In *Proceedings of Int'l Conf. Spoken Language Processing*.

Chanceaux, M., Guerin-Dugue, A., Lemaire, B., and Baccino, T. (2008). Towards a model of information seeking integrating visual semantic and memory maps. In *Proceedings of the 4th international cognitive vision workshop.*, pages 65–78.

Chun, M. and Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, 36:28–71.

Chun, M. and Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, 10:360–365.

Clarke, A., Coco, M., and Keller, F. (2013). The impact of attentional, linguistic and visual features during object naming. *Frontiers in Perception Science: Research Topic on Scene Understanding: Behavioral and computational perspectives*, under revision.

Coco, M. (2011). *Coordination of Vision and Language in Cross-Modal Referential Processing*. PhD thesis, School of Informatics (ILCC), University of Edinburgh.

Coco, M. and Keller, F. (2009). The impact of visual information on reference assignment in sentence production. In *Proceedings of CogSci*.

Condon, W. and Osgton, W. (1971). Speech and body motion synchrony of the speaker-hearer. In Horton, D. and Jenkins, J., editors, *The Perception of Language*, pages 150–184. Academic Press.

Cooper, R. (1974). Control of eye fixation by meaning of spoken language: New methodology for real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6:84–107.

Cristino, F., MathÃ´t, S., J., T., and Gilchrist, I. D. (2010). Scanmatch: A novel method for comparing fixation sequences. *Behaviour Research Methods*, 42:692–700.

Crocker, M., Knoeferle, P., and Mayberry, M. (2010). Situated sentence comprehension: The coordinated interplay account and a neurobehavioral model. *Brain and Language*, 112:189–201.

Dale, R. and Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, 56(3):391–430.

Davelaar, E. J., Goshen-Gottstein, Y., Haarmann, H. J., and Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigation of recency effects. *Psychological Review*, 112(1):3–42.

De Graef, P., Christiaens, D., and d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, 52:317–329.

Dewhurst, R., Nystrom, M., Jarodzka, H., Foulsham, T., Johansson, R., and Holmqvist, K. (2012). It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior Research Methods*, pages 1–22.

Dziemianko, M., Clarke, A., and Keller, F. (2011a). Towards object-based saliency. In *Proceedings of the 24th International Conference on Intelligent Robots and Systems (IROS)*.

Dziemianko, M., Clarke, A., and Keller, F. (2013). Object-based saliency as a predictor of attention in visual tasks. In *Proceedings of the 34nd Annual Conference of the Cognitive Science Society*.

Dziemianko, M., Coco, M., and Keller, F. (2011b). Increamental learning of target positions in visual search. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Dziemianko, M. and Keller, F. (2013). Memory modulated saliency: A computational model of the incremental learning of target locations in visual searchs. *Visual Cognition*, 21(3):277–305.

Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., and Oliva, A. (2009). Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 17(6/7):945–978.

Einhauser, W., Spain, M., and Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):1–26.

Ekman, P. (1979). About brows: Emotional and conversational signals. In *Human ethology: Claims and limits of a new discipline*.

Engbert, R., Nuthmann, A., Richter, E., and Kliegl, R. (2005). Swift: A dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777–813.

Eslami, S. and Williams, C. (2011). Factored shapes and appearances for parts-based object understanding. In *Proceedings of the British Machine Vision Conference*, pages 18.1–18.12. BMVA Press.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 22:861–874.

Findlay, J. and Gilchrist, I. (2001). *Visual attention: The active vision perspective*. Springer-Verlag, New York.

Fletcher-Watson, S., Findlay, J., Leekam, S., and Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, 37(4):571–583.

Foulsham, T., Dewhurst, R., Nystrom, M., Jarodzka, H., Johansson, R., Underwood, G., and Holmqvist, K. (2012). Comparing scanpaths during scene encoding and recognition: A multi-dimensional approach. *Journal of Eye Movement Research*, (5):1–14.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Gibson, B., Li, L., Skow, E., Brown, K., and Cooke, L. (2000). Searching for one versus two identical targets: When visual search has memory. *Psychological Science*, 11:324–327.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1:3–55.

Gomez, C. and Valls, A. (2009). A similarity measure for sequences of categorical data based on the ordering of common elements. *Lecture Notes in Computer Science*, 5285/2009:134–145.

Green, C. and Hummel, J. (2006). Familiar interacting object pairs are perceptually grouped. *Journal of Experimental Psychology: Human Perception and Performance*, 32:1107–1119.

Hare, M., Mcrae, K., and Elman, J. L. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, 48(2):281–303.

Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hayhoe, M., Shrivastava, A., Mruczek, R., and Pelz, J. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, 3:49–63.

Henderson, J. (2003). Human gaze control in real-world scene perception. *Trends in Cognitive Science*, 7:498–504.

Henderson, J., Brockmole, J., and Castelhano, M. (2007). Visual saliency does not account for eye-movements during visual search in real-world scenes. *Eye movements research: insights into mind and brain*.

Henderson, J., Malcolm, G., and Schandl, C. (2009). Searching in dark: Cognitive relevance drives attention in real-world scenes. *Psycholomic Bulletin & Review*, 16:850–856.

Henderson, J., Weeks Jr., P., and Hollingworth, A. (1999). Effects of semantic consistency on eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 25:210–228.

Hollingworth, A. (2006). Visual memory for natural scenes: Evidence from change detection and visual search. *Visual Cognition*, 14:781–807.

Hollingworth, A., Williams, C. C., and Henderson, J. M. (2001). To see and remember: Visually specific information is retained in memory from previously attended objects in natural scenes. *Psychonomic Bulletin & Review*, 8:761–768.

Honda, K. (2000). Interactions between vowel articulation and F0 control. In *In Proceedings of Linguistics and Phonetics: Item Order in Language and Speech (LP?98)*.

Horowitz, T. and Wolfe, J. (1998). Visual search has no memory. *Nature*, 394(6):575–577.

Hwang, A., Wang, H., and Pomplun, M. (2011). Semantic guidance of eye movements during real-world scene inspection. *Vision Research*, 51(10):1192–1205.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12:1093–1123.

Itti, L., Koch, C., and 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.

Jarodzka, H., Holmqvist, K., and Nystrom, M. A. (2010). Vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 211–218. ACM.

Johnston, W. and Dark, V. (1986). Annual review of psychology. *Selective Attention*, 37:43–75.

Klein, D. and Frintrop, S. (2012). Salient pattern detection using w2 on multivariate normal distributions. In *Joint Pattern Recognition Symposium of the German Association for Pattern Recognition (DAGM) and the Austrian Association for Pattern Recognition (OAGM) (DAGM-OAGM)*.

Klein, R. (1988). Inhibitory tagging system facilitates visual search. *Nature*, 334:430–431.

Klein, R. and MacInnes, W. J. (1999). Inhibition of return is a foraging facilitator in visual search. *Psychological Science*, 10:346–352.

Knoeferle, P. and Crocker, M. (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*, (30):481–529.

Koostra, G., Nedereen, A., and De Boer, B. (2008). Paying attention to symmetry. In *Proceedings of British Machine Vision Conference*.

Krantz, J. (2012). *Experiencing Sensation and Perception*. Pearson Education.

Kukona, A. and Tabor, W. (2011). Impulse processing: A dynamical systems model of incremental eye movements in the visual world paradigm. *Cognitive Science*, 35.

LaBerge, D. (1983). Spatial extent of attention to letters and words. *Journal of Experimental Psychology: Human Perceprion and Performance*, 3(9):371–379.

Lafferty, J., McCallum, A., and Pereira, F. (2000). Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *In Proceedings International Conference on Machine Learning*.

Land, M. and Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, 41:3559–3565.

Lee, S. P., Badler, J., and Badler, N. (2002). Eyes alive. In *SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 637–644, New York, NY, USA. ACM Press.

Leek, E., Cristino, F., Conlan, L., Rodriguez, E., and Johnston, S. (2012). Fixational eye movement patterns during object shape perception: A fixation preference for concave surface curvature minima. *Journal of Vision*, in press.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., and Shum, H. (2011). Learning to detect salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353 – 367.

Liu, X., Zhao, Y., Pi, X., Liang, L., and Nefian, A. (2002). Audio-visual continuous speech recognition using a coupled hidden markov model. In *IEEE International Conference on Spoken Language Processing*.

Loftus, G. (1985). Picture perception: Effects of luminance level on available information and information-extraction rate. *Journal of Experimental Psychology: General*, 114:342–356.

Loftus, G., Kaufman, L., Nishimoto, T., and Ruthruff, E. (1992). *Why it's annoying to look at slides with the room lights still on: Effects of visual degradation on perceptual processing and long-term visual memory*, pages 203–226. Springer-Verlag.

Lyngso, R., Pedersen, C., and Nielsen, H. (1999). Metrics and similarity measures for hidden markov models. In *Proc Int Conf Intell Syst Mol Biol*.

Madsen, A., Larson, A., Loschky, L., and Rebello, N. S. (2012). Using scanmatch scores to understand differences in eye movements between correct and incorrect solvers on physics problems. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, pages 193–196, New York, NY, USA. ACM.

Maljkovic, V. and Martini, P. (2005). Implicit short-term memory and event frequency effects in visual search. *Vision Research*, 45(21):2831–2846.

Marwan, N. and Kurths, J. (2002). Nonlinear analysis of bivariate data with cross recurrence plots. *Physics Letters A*, 302(5–6):299–307.

Mayberry, M., Crocker, M., and Knoeferle, P. (2005). A connectionist model of sentence comprehension in visual worlds. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

Mayberry, M., Crocker, M., and Knoeferle, P. (2009). Learning to attend: A connectionist model of situated language comprehension. *Cognitive Science*, 33(1):449–496.

McPeek, R. M., Maljkovic, V., and Nakayama, K. (1999). Saccades require focal attention and are facilitated by a short-term memory system. *Vision Research*, 39(8):1555–1566.

Metropolis, N.and Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092.

Molina, A. and Pla, F. (2002). Shallow parsing using specialized hmms. *Journal of Machine Learning Research*, 2:595–613.

Nefian, A., Liang, L., Pi, X., Mao, C., and Murphy, K. (2002a). A coupled hmm for audio-visual speech recognition. In *International Conference on Acoustics, Speech and Signal Processing*.

Nefian, A., Liang, L., Pi, X., and Murphy, K. (2002b). Dynamic bayesian networks for audio-visual speech recognition. *Journal of Applied Signal Processing*.

Noton, D. and Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, 171(3968):308–311.

Nuthmann, A. and Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, 20(10(8)):1–19.

Oliva, A. and Schyns, P. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41:176–210.

Oliva, A. and Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527.

Otero-Millan, J., Troncoso, X., Macknik, S., Serrano-Pedraza, I., and Martinez-Conde, S. (2008). Saccades and microsaccades during visual fixation, exploration, and

search: Foundations for a common saccadic generator. *Journal of Vision*, 14(8):1–18.

Pate, J. and Goldwater, S. (2011). Unsupervised syntactic chunking with acoustic cues: Computational models for prosodic bootstrapping. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*.

Pearl, J. (1982). Reverend bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the Second National Conference on Artificial Intelligence (AAAI-82)*, pages 133–136.

Pelz, J. and Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research*, 41:3587–3596.

Peters, R., Iyer, A., Itti, L., and Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, 45:2397–2416.

Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research*, 46:1886–1900.

Posner, M. (1978). *Chronometric Explorations of Mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Posner, M. (1980). Orienting of attention. the 7th sir f.c. barlett lecture. *Quaterly Journal of Experimental Psychology*, 32:3–25.

Potter, M. C. (1975). Meaning in visual scenes. *Science*, 187:965–966.

Potter, M. C., Staub, A., and O'Connor, D. H. (2004). Pictorial and conceptual representation of glimpsed pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 30:478–489.

Prasov, Z. and Chai, J. (2008). What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *IUI*, pages 20–29.

Prasov, Z., Chai, J., and Jeong, H. (2007). Eye gaze for attention prediction in multimodal human-machine conversation. In *Interaction Challenges for Intelligent Assistants*, pages 102–110.

Qu, S. and Chai, J. (2008). Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In *EMNLP*, pages 244–253.

Qu, S. and Chai, J. (2009). The role of interactivity in human-machine conversation for automatic word acquisition. In *SIGDIAL Conference 2009*, pages 188–195.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124:372–422.

Reichle, E., Rayner, K., and Pollatsek, A. (2003). The e-z reader model of eye-mevement control in reading: Coparisons to othe models. *Behavioral and Brain Sciences*, 26:445–526.

Renninger, L. and Malik, J. (2004). When is scene identification just texture recognition? *Vision Research*, 44:2301–2311.

Rensink, R. (2000a). The dynamic representation of scenes. *Vision Cognition*, 1/2/3(7):17–42.

Rensink, R. (2000b). Seeing, sensing, and scrutinizing. *Vision Research*, 10-12(40):1469–1487.

Richardson, D., Dale, R., and Kirkham, N. (2007). The art of conversation is co-ordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science*, 18(5):407 – 413.

Roy, D. and Mukherjee, N. (2005). Towards situated speech understanding: Visual context priming of language models. *Computer Speech and Language*, 19(2):227–248.

Schauerte, B., Kuhn, B., and Kroschel, K. (2012). Multimodal saliency-based attention for object-based scene analysis. In *The 3rd International Conference on Appearance (Predicting Perceptions 2012)*.

Schmidt, J. and Zielinsky, G. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quaterly Journal of Experimental Psychology*, 62:1904–1914.

Schyns, P. and Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, 5:195–200.

Sedivy, J., Tanenhaus, M., Chambers, C., and Carlson, G. (1999). Achieving incremental interpretation through contextual representation: Evidence from the processing of adjectives. *Cognition*, 71:109–147.

Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL 03*, pages 213–220.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Shore, D. and Klein, R. (2000). On the manifestations of memory in visual search. *Spatial Vision*, 14:59–75.

Simola, J., Salojarvi, J., and Kojo, I. (2008). Using hidden markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research*, 9:237–251.

Snedeker, J. and Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, 48:103–130.

Soding, J. (2005). Protein homology detection by hmm-hmm comparison. *Bioinformatics*, 22(7):951–960.

Spain, M. and Perona, P. (2011). Measuring and predicting object importance. *International Journal of Computer Vision*, 91:59–76.

Spivey, M. and Dale, R. (2006). Continuous temporal dynamics in cognition. *Current Directions in Psychological Science*, 15:2007–2011.

Spivey-Knowlton, M., Tanenhaus, M., Eberhard, K., and Sedivy, J. (2002). Eye movements and spoken language comprehension: Effects of syntactic context on syntactic ambiguity resolution. *Cognitive Psychology*, 4(45):447–481.

SR Research (2002). Eyelink II user manual.

Takeda, Y. and Yagi, A. (2000). Inhibitory tagging in visual search can be found if search stimuli remain visible. *Perception and Psychophysics*, 62:927–934.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Tatler, B. W. (2007). The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):1–17.

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

Torralba, A., Oliva, A., Castelhano, M., and Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, 113:766–786.

Underwood, G. and Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology*, 59(11):1931–1949.

Vo, M. L.-H. and Henderson, J. M. (2009). Does gravity matter? effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, 9(3):15.

Vo, M. L.-H. and Henderson, J. M. (2010). The time course of initial scene processing. *Visual Cognition*, 18(1):148–152.

Walther, D. and Koch, C. (2006). Modelling attention to salient proto-objects. *Neural Networks*, 19:1395–1407.

Wolfe, J. (1998). Visual memory: What do you know about what you saw? *Current Biology*, 8:303–304.

Wolfe, J. M. (1999). Inattentional amnesia. In Coltheart, V., editor, *Fleeting memories*, pages 71–94, Cambridge, MA. MIT Press.

Wolfe, J. M. and Bennett, S. C. (1997). Preattentive object files: Shapeless bundles of basic features. *Vision Research*, 37:25–44.

Wolfe, J. M., Klempen, N., and Dahlen, K. (2000). Postattentive vision. *Journal of Experimental Psychology: Human Perception and Performance*, 26:693–716.

Woodman, G. F. and Chun, M. M. (2006). The role of working memory and long-term memory in visual search. *Visual Cognition*, 14:808–830.

Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum Press.

Yehia, H., Kuratate, T., and Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. In *Journal of Phonetics*.

Young, R. and O'Shea, T. (1981). Errors in children's subtration. *Cognitive Science*, 5(2):153–177.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). The htk book.

Yu, Y., Mann, G., and Gosine, R. (2010). An object-based visual attention model for robotic applications. *IEEE Transactions on Systems, Man, and Cybernetics: Cybernetics*, 5(40):1398–1412.

Zelinsky, G. (2008). A theory of eye movements during target acquisition. *Psychological Review*, 115:419–433.

Zelinsky, G. (2012). Tam: Explaining off-object fixations and central fixation tendencies as effects of population averaging during search. *Visual Cognition*, 20(4-5):515–545.

Zelinsky, G. and Schmidt, J. (2009). An effect of referential scene constraint on search implies scene segmentation. *Visual Cognition*, 17:1004–1028.

Zhang, L., Tong, M., Marks, T., Shan, H., and Cottrell, G. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 32(8(7)):1–20.