
# A computational model of eye guidance, searching for text in real scene images

A dissertation submitted by **Antonio Clavelli** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor en Informàtica**.

Bellaterra, July 2014

| | |
|---|---|
| Director | **Dr. Dimosthenis Karatzas**<br>Centre de Visió per Computador<br>Universitat Autònoma de Barcelona |
| Co-director | **Prof. Giuseppe Boccignone**<br>Dipartimento di informatica<br>Università degli studi di Milano |
| Tutor | **Prof. Josep Lladós**<br>Centre de Visió per Computador<br>Universitat Autònoma de Barcelona |
| Thesis Committe | **Prof. Angelo Marcelli**<br>Dipartimento DIEM<br>Università degli studi di Salerno<br>**Prof. Sophie Wuerger**<br>Department of Psychological Sciences<br>University of Liverpool<br>**Prof. Alejandro Parraga**<br>Centre de Visió per Computador<br>Universitat Autònoma de Barcelona<br>**Dr. Laura Dempere**<br>Department of Information and Communication Technologies<br>Universitat Pompeu Fabra<br>**Prof. Maria Vanrell**<br>Centre de Visió per Computador<br>Universitat Autònoma de Barcelona |


This document was typeset by the author using LaTeX 2$\varepsilon$.

The research described in this book was carried out at the Computer Vision Center, Universitat Autònoma de Barcelona.

Copyright © 2014 by Antonio Clavelli. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the author.

ISBN 978-84-940902-6-4-ISBN-NUMBER

Printed by Ediciones Gráficas Rey, S.L.

*to my wife and family*

*"Believe you can and you're halfway there."*

— Theodore Roosevelt

# Acknowledgements

I'm extremely thankful to the CVC for having given me the opportunity to conduct my PhD research. It has been an extremely powerful experience made extremely rich by the exceptional people I met. I would like to express my special appreciation and thanks to:

- Dr. Dimosthenis Karatzas, for having guided me through the research journey since the first day with patience and wisdom. We managed to solve many challenges and it has been a pleasure working with you. Thanks for making this a rich and empowering experience.
- Prof. Josep Lladós for welcoming me to CVC in the role of research centre director and for tutoring my thesis. Thanks for having removed many obstacles allowing me to be productive.
- Prof. Giuseppe Boccignone for having guided me through the cognitive aspects of vision. I learned a lot from you. Thanks for having inspired me passion for research and thoughts.
- Prof. Mario Ferraro for his insightful comments and feedback on the manuscript and articles. Thanks a lot.
- The Agency for Management of University and Research Grants, AGAUR, to have supported my research with a three years fellowship number 2009FIB00020.
- All my friends and colleagues, both old and new members from the CVC and the University of Milan. A big hug goes to my earliest colleagues who I met at the University of Salerno. Thanks to all of you for the thoughtful conversations, great tips and enjoyable leisure time.
- My wife and family to be supportive and trustful. Thanks for your encouragement, there is no way that I would have been able to make it without your love and support.

# Abstract

Searching for text objects in real scene images is an open problem and a very active computer vision research area. A large number of methods have been proposed tackling the text search as extension of the ones from the document analysis field or inspired by general purpose object detection methods. However the general problem of object search in real scene images remains an extremely challenging problem due to the huge variability in object appearance. This thesis builds on top of the most recent findings in the visual attention literature presenting a novel computational model of eye guidance aiming to better describe text object search in real scene images.

First are presented the relevant state-of-the-art results from the visual attention literature regarding eye movements and visual search. Relevant models of attention are discussed and integrated with recent observations on the role of top-down constraints and the emerging need for a layered model of attention in which saliency is not the only factor guiding attention. Visual attention is then explained by the interaction of several modulating factors, such as objects, value, plans and saliency.

Then we introduce our probabilistic formulation of attention deployment in real scene. The model is based on the rationale that oculomotor control depends on two interacting but distinct processes: an attentional process that assigns value to the sources of information and motor process that flexibly links information with action. In such framework, the choice of where to look next is task-dependent and oriented to classes of objects embedded within pictures of complex scenes. The dependence on task is taken into account by exploiting the value and the reward of gazing at certain image patches or proto-objects that provide a sparse representation of the scene objects.

In the experimental section the model is tested in laboratory condition, comparing model simulations with data from eye tracking experiments. The comparison is qualitative in terms of observable scan paths and quantitative in terms of statistical similarity of gaze shift amplitude. Experiments are performed using eye tracking data from both a publicly available dataset of face and text and from newly performed eye-tracking experiments on a dataset of street view pictures containing text.

The last part of this thesis is dedicated to study the extent to which the proposed model can account for human eye movements in a low constrained setting. We used a mobile eye tracking device and an ad-hoc developed methodology to compare model simulated eye data with the human eye data from mobile eye tracking recordings. Such setting allow to test the model in an incomplete visual information condition, reproducing a close to real-life search task.

# Resum

La cerca d'objectes de text en imatges d'escena reals és un problema obert i una àrea de cerca molt activa la visió per computador. S'han proposat un gran nombre de mètodes basats en l'extensió dels mètodes des de l'anàlisi de documents o inspirat en mètodes de detecció d'objectes. No obstant això, el problema de la cerca d'objectes en imatges d'escena reals segueix sent un problema extremadament difícil a causa de la gran variabilitat en l'aparença dels objectes. Aquesta tesi es basa en els més recents troballes en la literatura de l'atenció visual, introduint un nou model computacional de visió guiada que apunta descriure la cerca de text en imatges d'escenes reals.

En primer lloc es presenten els resultats més pertinents de la literatura científica en relació amb l'atenció visual, els moviments oculars i la cerca visual. Els més rellevants models d'atenció són discutits i integrats amb recents observacions sobre la funció dels anomenats 'top-down constraints' i l'emergent necessitat d'un model estratificat d'atenció en què la saliència no és l'únic factor guia d'atenció. L'atenció visual s'explica per la interacció de diversos factors moduladors, com ara objectes, valor, plans i saliència.

S'introdueix la nostra formulació probabilística dels mecanismes d'atenció en escenes reals per a la tasca de cerca d'objectes. El model es basa en l'argument que el desplegament d'atenció depèn de dos processos diferents però interactuants: un procés d'atenció que assigna valor a les fonts d'informació i un procés motor que uneix flexiblement informació amb l'acció. En aquest marc, l'elecció d'on buscar la propera tasca és dependent i orientada a les classes d'objectes incrustats en imatges d'escenes reals. La dependència de la tasca es té en compte en explotar el valor i la recompensa de contemplar certes parts o proto-objectes de la imatge que proporcionen una esclarissada representació dels objectes en l'escena.

A la secció experimental prova el model en condicions de laboratori, comparant les simulacions del model amb dades d'experiments de eye tracking. La comparació és qualitativa en termes de trajectòries d'exploració i quantitativa, en termes de similitud estadística de l'amplitud de moviments oculars. Els experiments s'han realitzat amb dades de eye tracking tant d'un conjunt de dades públic de rostre humans i text, tant amb un nou conjunt de dades de eye tracking i d'imatges urbanes amb text.

L'última part d'aquesta tesi es dedica a estudiar en quina mesura el model proposat pot respondre del desplegament d'atenció en un entorn complex. S'ha utilitzat un dispositiu mòbil de eye tracking i una metodologia desenvolupada específicament per comparar les dades simulades amb les dades gravades de eye tracking. Tal configuració permet posar a prova el model en la tasca de cerca de text molt semblant a una cerca real, en la condició d'informació visual incompleta.

# Resumen

La búsqueda de objetos de texto en imágenes de escena reales es un problema abierto y un área de investigación muy activa la visión por computador. Se han propuesto un gran número de métodos basados en la extensión de los métodos desde el análisis de documentos o inspirado en métodos de detección de objetos. Sin embargo, el problema de la búsqueda de objetos en imágenes de escena reales sigue siendo un problema extremadamente difícil debido a la gran variabilidad en la apariencia de los objetos. Esta tesis se basa en los más recientes hallazgos en la literatura de la atención visual, introduciendo un nuevo modelo computacional de visión guiada que apunta a describir la búsqueda de texto en imágenes de escenas reales.

En primer lugar se presentan los resultados mas pertinentes de la literatura científica en relación con la atención visual, los movimientos oculares y la búsqueda visual. Los mas relevantes modelos de atención son discutidos e integrados con recientes observaciones sobre la función de los denominados 'top-down constraints' y la emergente necesidad de un modelo estratificado de atención en el que la saliencia no es el único factor guía de atención. La atención visual se explica por la interacción de varios factores moduladores, tales como objetos, valor, planes y saliencia.

Se introduce nuestra formulación probabilística de los mecanismos de atención en escenas reales para la tarea de búsqueda de objetos. El modelo se basa en el argumento de que el despliegue de atención depende de dos procesos distintos pero interactuantes: un proceso de atención que asigna valor a las fuentes de información y un proceso motor que une flexiblemente información con la acción. En ese marco, la elección de dónde buscar la próxima tarea es dependiente y orientada a las clases de objetos incrustados en imágenes de escenas reales. La dependencia de la tarea se tiene en cuenta al explotar el valor y la recompensa de contemplar ciertas partes o proto-objetos de la imagen que proporcionan una rala representación de los objetos en la escena.

En la sección experimental se prueba el modelo en condiciones de laboratorio, comparando las simulaciones del modelo con datos de experimentos de eye tracking. La comparación es cualitativa en términos de trayectorias de exploración y cuantitativa, en términos de similitud estadística de la amplitud de movimientos oculares. Los experimentos se han realizado con datos de eye tracking tanto de un conjunto de datos públicos de rostros humanos y texto, tanto con un nuevo conjunto de datos de eye tracking y de imágenes urbanas con texto.

La última parte de esta tesis se dedica a estudiar en qué medida el modelo propuesto puede responder del despliegue de atención en un entorno complejo. Se ha utilizado un dispositivo móvil de eye tracking y una metodología desarrollada específicamente para comparar los datos simulados con los datos grabados de eye tracking. Tal configuración permite poner a prueba el modelo en la tarea de búsqueda de texto muy parecida a una búsqueda real, en la condición de información visual incompleta.

# Contents

# List of Figures

# Chapter 1

# Introduction

Searching for objects in real scene images is known to be a difficult problem due to the huge variability of object appearance. The majority of computer vision approaches commonly address the problem of object detection by means of class-specific trained object detectors [136, 16]. However, begin to appear, novel computer vision approaches aiming to detect objects in the large, generating possible object locations for use in object recognition [118].

Parallel to the research in computer vision, the fields of cognitive science and psychology have made considerable progress addressing the principles of *vision*, aiming to explain how the human visual system accomplish *visual search task*. Abstract models of attention attempt to trace the ongoing cognitive process and describe them at the functional and biological level. The study of eye movements proved to be of great value as providing observable fact deeply related to the cognitive process [88] and starting with the pioneering work by Koch and Ullman in 1985, several saliency based computational models of attention have been proposed in the attempt to describe the eye movements and attention deployment mechanism.

A basic aspect of the human visual system is its limited visual acuity that seems to be a prerequisite for the efficient and effective navigation of the surrounding world. The foveal area of the retina extends about two degrees across the center of the eye and achieves the highest visual acuity in a vary narrow field of view. The nonuniform resolution of the human eye sensor gives rise to the necessity of eye movements as a mechanism to move the fovea on the part of the scene we need to see clearly.

Recent findings on the so called phenomenon of change blindness, highlighted special experimental settings in which humans fail at the perceptual task of detecting changes in the scene. As such it tells us about the strategies employed by the visual system to quickly and accurately process the huge amount of information coming from the outer world [91]. Change blindness is a surprising and counter-intuitive finding, as the daily visual experience appears as a detailed and complete representation of the world. Human vision is far different than a photographic snapshot of the world. Investigation in the visual attention literature have clearly proved that vision is the result of a dynamic process based on a dynamic representation of the perceived world [90, 101]. Building on top of these findings, this thesis explore new ways computers

can be instructed to tackle complex visual search tasks, and introduces a novel computational model of eye guidance to describe the object search under a specific search task.

## 1.1 Motivations

Attention based mechanisms are naturally embedded in the human visual system as a mean for realizing an active vision process. In this perspective, first recall that, although we deal with a perceptual task (e.g. a text localization/detection), our active vision approach diverges from the traditional concept in computer vision that: sensation, perception and cognition are isolated processes previous to actuation (passive vision). Under the passive vision paradigm the perceptual system is limited to operate using the raw data captured by the sensors "as is". In active vision, we begin not with a sensory stimulus, but with a sensorimotor coordination. It is the movement which is primary, and the sensation which is secondary, the movement of the body, head, and eye muscles determining the quality of what is experienced. Therefore, the active observer does not obtain information by plain observation, but also by interaction and selection of stimuli so to gain control of what to see and how to see it.

This point can be further described in terms of *foraging metaphor*. In a foraging metaphor [10, 129] the eye is a forager moving across the visual landscape and feeding on valuable information. The forager, moment to moment, is confronted with the choice between "feed", that is, performing local intensive exploration of the landscape, or to "fly" by making an extensive relocation toward over the landscape. This choice, in turn, entails the decision of whether to stay longer or to fly that is usually based on incomplete information.

**When the active vision approach may be of vital importance for the localization/detection task?** We focus here on the *Visibility / Discriminability* space, a simplified representation as described in Fig. 1.1 that can be useful to provide some hints on the issues discussed by Eckstein under the question [31].: What limits visual search performance?

**Visibility** Visibility is related to the location of the potential target with respect to the current focus of attention and related field of view (FOV) of the observer. In a full visibility condition all targets are equally visible within the foveal region and the observer just needs to perform a detection procedure. This is the case corresponding to classic machine vision and pattern recognition procedure.

Visibility decreases when targets become located in the extra-foveal field of view, and decreases even further in the extreme case when the targets are out of the FOV: consider for instance yourself walking down an unknown street and looking for an hotel: you will move your eyes, move your head, or even stop and turn your back with your body to be sure you have not previously missed the target. That is actually searching in the wild. The visibility is also a well studied problem in the visual attention literature. From the searcher / forager standpoint full visibility corresponds to perfect information of target location, thus the foraging problem is just to perform an optimal tour while foraging.

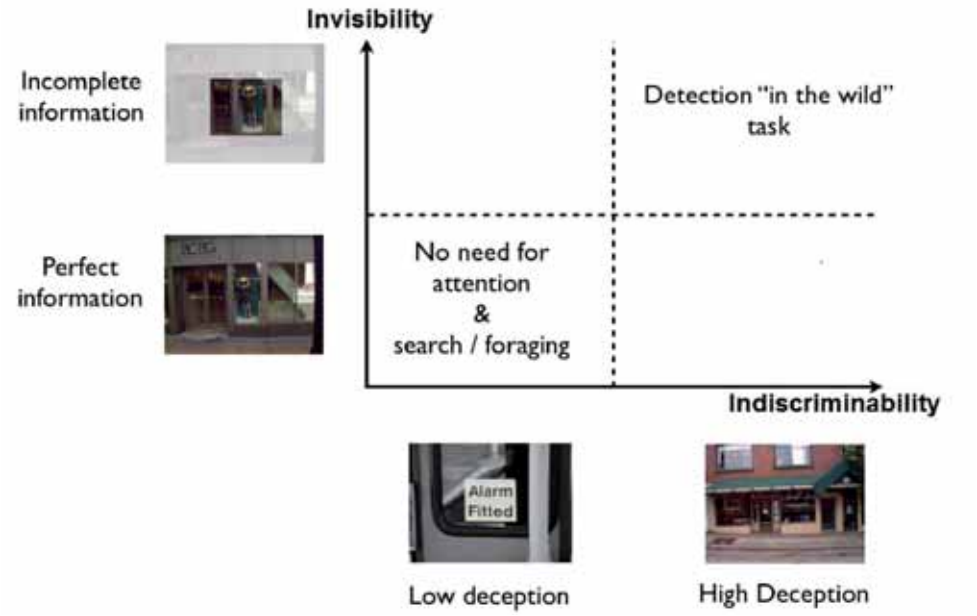



**Figure 1.1:** The problem of the searcher: Visibility vs. Discriminability in detection task

**Discriminability** Discriminability relates to how difficult is to effectively compute the likelihood of an object's location. In high discriminability condition object's feature are fully informative, and a simple detection procedure is needed; low discriminability entails that features we can extract are are not really distinctive, and in turn this implies the adoption of a more complex detection procedure. From the searcher / forager standpoint discriminability is related to the deceitfulness of the target/prey. Recalling Wang and Pomplun's experiments, observers are attracted by text, but we can deceive the observer, undermining his recall/precision performance by inserting spurious patches that may have spatial-frequency text-like characteristics yet they are not text.

Thus, if considering the Visibility / Discriminability space in the Figure 1.1 one can feel quite comfortable in stating that in the full Visibility / Discriminability condition we certainly do not need to resort to the active paradigm. On the other extreme, in very-low Visibility / Discriminability, we are forced to exploit an active approach; in terms of foraging it means that in null visibility condition the forager must rely on a stochastic search. Other cases are probably in between these two.

## 1.2 Objective of this work

This thesis is part of a larger research aiming to develop state-of-the-art human inspired text localization algorithms able to cope with condition of incomplete visual information and high deception, as in a real-life 360 degree search. To such end the attention mechanisms are implemented as an active vision mechanism to explore and detect potential objects of interest. The goals of this thesis are:

1. investigate the mechanism behind the attention allocation in real scene images and provide a plausible computational model able to describe how some top-down constrains can interact with stimulus driven saliency. We are especially interested in the task of object search and specifically text-object search in street view images.

2. describe statistical properties of model generated and eye tracked gaze shifts as closely as possible, including inter-individual scan path variability.

3. measure the extent to which the model of attention can account for human eye movements in a close to real life experimental conditions, accounting for incomplete visual information condition.

## 1.3 Contributions

1. Proposed a novel computational model of eye guidance extending the existing models of attention to include a multilevel description of the scene at the saliency, object and value level. The problem of text search is then tackled as a foraging problem in which the 'foraging eye' moves across a multilevel description of the scene, encoding salience and task-dependent information oriented to classes of objects present in the scene. The model, tested in laboratory experiments, proved to account well the statistical properties of gaze shifts.

2. Studied to what extent the model can account for human eye movements in condition of high deception and incomplete information. Developed an ad-hoc methodology to compare model simulated eye data with the human eye data from mobile eye tracking recordings.

3. Created two new eye tracking datasets:

   (a) an eye tracking dataset for text detection in real scene images to study the influence of the task on attention deployment. Eye tracking data are recorded under a free viewing and a look for text experimental condition, using a traditional desktop eye tracker.

   (b) An eye tracking dataset for text search in outdoor setting to investigate eye movements under a condition of incomplete visual information in a real-life 360 degree search. Data are collected by making use of a mobile eye tracker to allow low constrained experimental conditions and freedom of movements for the subjects under test.

## 1.4 Organization

The thesis is organized as follows. In Chapter 2 we will first review the state-of-the-art on attention modeling. The content is a logical reorganization of the relevant contributions made in the visual attention literature regarding eye movements and visual search, constituting the backbone of our approach to attention modeling. In Chapter 3 we will describe the proposed computational model of attentional eye guidance and a rigorous experimental evaluation on the gaze shift amplitude distributions will be carried out in Chapter 4. In Chapter 5 we will show to what extent the proposed model of attention can mimic observer's oculomotor behaviour in outdoor settings. Chapter 6 will conclude this thesis by summarizing the main contributions of the work presented hereafter. It will also point out the limits of this model and the future research directions that this thesis may open. Finally, Appendixes A will present additional material for the interested readers about foraging models and Lévy flights.

# Chapter 2

# Eye Movements, Attention and Visual search

## 2.1 Eye Movements

The visual field can be divided in three regions: the foveal, parafoveal and peripheral. The fovea is the central area of the retina, extending about two degrees across the center of the eye, and achieves the best visual acuity. Moving from the fovea toward the periphery, the visual acuity drops quickly as shown in figure 2.2. The parafovea surrounds the fovea and is poorer in visual acuity although it extends out to ten degrees off-center, the periphery is the largest region and the one with the lowest acuity. Such limited visual acuity give rise to the necessity of eye movements to bring the fovea on the part of the scene we need to see clearly. Several kinds of movements have been observed although their role is not completely understood [63].

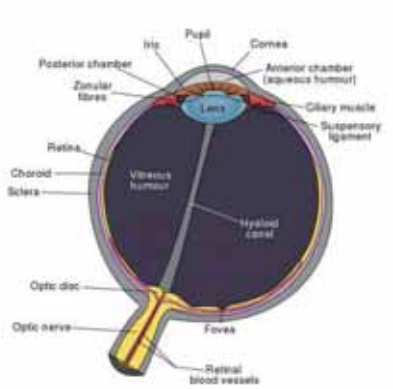



**Figure 2.1:** Schematic diagram of the human eye, with the fovea at the bottom. It shows a horizontal section through the right eye. From Wikipedia.

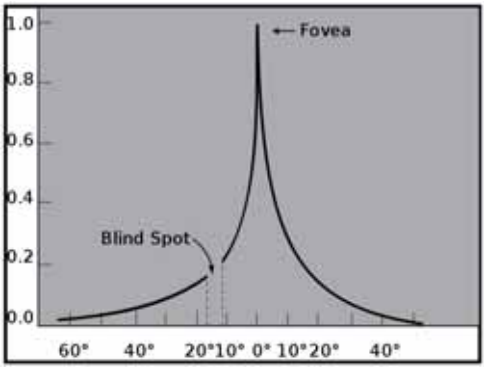



**Figure 2.2:** The relative acuity of the human eye quickly drops as as moving from the fovea toward the periphery. Image from Wikipedia.

Saccades are the fastest eye movements, they allow the jumping to different portions of the scene. During saccades vision is inhibited and no visual information is gained. The phenomenon goes under the name of saccadic suppression [67] and it is completely automatic and volition independent. From a dynamical point of view saccades differ from other eye movements by their ballistic nature. The velocity rapidly rises during the saccade to a maximum that occurs slightly before the midpoint of the movement and then drops at a slightly slower rate until the target location is reached. According to previous experiments reported by Rayner [88] the velocity of the saccade is a monotonic function of how far the eyes move. A 2 degrees saccade typical of reading takes around 30 ms, whereas a 5 degrees saccade, typical of scene perception, takes around 40-50 ms.

Pursuits eye movements are the movements initiated to follow a moving target. They are an active response to the stimulus, able to synchronize the fovea's speed to the object's cues such as speed, attention, expectation. Pursuits are probably completely involuntary movements as it does not seems to be possible initiate pursuits without a proper stimulus. Pursuits movements are slower than saccades and, can be interlaced by some saccades to catch up with the target when its moving too fast. Recent studies have begun to blur the classical line between pursuits and saccades [63] and evidence has been gained for a close coupling between the control of selection for pursuits and saccades.

Fixations relate to an almost still eye condition, typically lasting about 200-300 ms. In order to maintain stable the image on the retina during fixations a slow control is continuously performed and there is a broad agreement that image motion on retina is crucial for vision and if too much motion degrades resolution, too little image motion may lead to image fading [63]. Microsaccades are saccades during fixations of a very small length (15 min arc) that do not seem to be due to image stabilization nor due to ideal generation of useful image motion. Recent research suggests that their role seems to be more related to the fovea repositioning on close details, that is exactly the role played by saccades on a bigger scale [63].

Some other movements, such as Vergence and Vestibular, are instead compensation

movements useful to keep stable the image on the retina. Vergence are inward eye movements occurring when, in order to fixate on a close object, we move our eyes toward each other. The Vestibular are instead rotational eye movements occurring to compensate head movements in order to maintain the same direction of vision.

### 2.1.1 Eye movements and perception

Thinking of perception as the task of attending visual information, it is clear that performance depends on the reliability of the available signal. Studies on the contrast sensitivity and visibility models provided a sound basis to measure the information gained by looking to specific locations. Work by Geisler [70] supports the thesis that humans select fixation locations in order to maximize the information gain more than to account for the location having the highest probability to be attended. Observers had to search for small Gabor targets in the midst of pink random noise, Geisler found that human performance closely matched the performance of an ideal Bayesian observer using just the knowledge about the visibility map to decide where to look next. [70]

Although the evidence provided by Geisler does not directly demonstrate that humans follow the exact computations of the ideal observer, it remains that eye movements and perception are strongly coupled. A single object starting to move is a typical example of signal able to trigger a saccade and consequent smooth pursuits to follow the movement. As such the moving object has to be visible enough to stand out from a noisy environment.

Other studies by Araujo et. al. [1] proposed that the next saccade generation might not be optimal in terms of information maximization. Authors used a simple, two-location visual search task and found that saccadic patterns were not much influenced by the probability of finding the target, instead a stronger correlation was observed with spatial and distance stimulus distribution. Their findings support the thesis that eye movements have a built-in preference to minimize the effort of cognitive and attentional load in saccadic planning and observed a preference to make saccades to nearby locations. Than looking at these findings it seems that the eye movements do not follow any simple strategy and eye movements cannot be explained only on the basis of perceptual stimulus. More complex mechanism are involved and attention is an often used concept to describe the eye movement guidance.

### 2.1.2 Eye movements and attention

The relationship between eye movements and attention has been extensively investigated. In reading task or any visual search tasks, the covert and the overt attention (eye location) are tightly linked [17] and the eye movements are commonly used as a measurements of covert attention during complex cognitive processing tasks [89]. Although it seems we can shift attention independently from the eye movements and, as a matter of fact, the planning of an eye movement is thought to be preceded by a shift of covert attention to the target location before the actual movement is deployed [88, 46]. What actually attracts attention, the role of covert attention and in general how eye movements are planned is a wide research area and the debate is open. Back

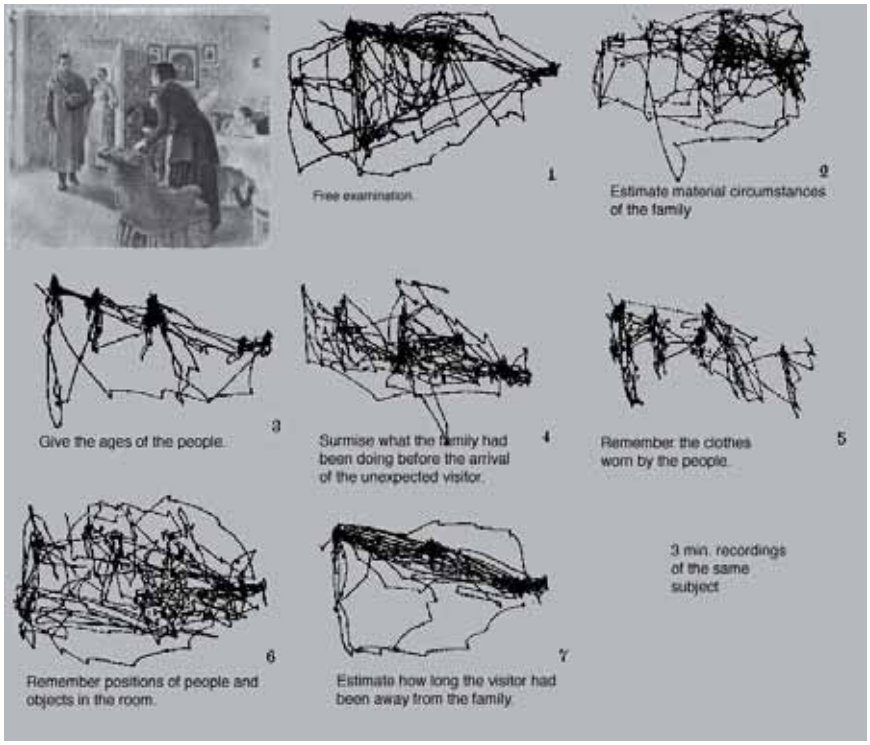



**Figure 2.3:** Yarbus. Studies on saccadic eye movements and subject's scanpath under several different task conditions.

in the seventies the associationist theories asserted that the experience of the whole is built by combining elementary sensations, on the other hand the Gestalt psychologists supported synthetic theories claiming that the whole proceeds its parts.

It has been thought that features come first in perception. The feature integration theory by Treisman [117] largely influenced research. An early parallel integration of features such as color, orientation, spatial frequency, brightness and direction of movements provides an initial coding of the scene. Then a spotlight of focused attention integrates the features by serially directing attention to the locations creating an unitary percept.

Yarbus [132] first performed important studies of saccadic eye movements showing that the subject's scanpath is highly influenced by the task that the observer has to perform. Some scanpaths from the Repin's painting *The Unexpected Visitor* are shown in Figure 2.3. Note the difference between plot 2 and plot 3 relatively to the task *estimate material circumstances of the family* and the task *give the age to the people*. It shows that subjects use scanning patterns that are quite dependent to the task at hand and fixations are highly linked to regions of interest.

At this regard the selection of a fixation points appears to be driven by a bottom-up process accounting for the saliency of image features and a top-down process producing a task-dependent and volition-controlled allocation of attention.

## 2.2 Vision and scene representation

One of the open problems in vision is about the kind of representation of the world might be retained by the brain. The problem of the brain's representation of the scene is linked to the capability of predicting the interesting fixation locations, based on a partial knowledge of the scene. In the observable facts vision is made up by frequent saccadic relocation while the retinal information has high resolution in the small central fovea, rapidly degrading radially towards the very low resolution peripheral vision, as shown in Figure 2.2.

An example used by Ballard [96] to describe the 'Vision problem' is the task of looking for a cup while preparing a cup of coffee. If we assume that the cup is not placed in the fovea, more likely be placed somewhere in the peripheral region of the retina, that the peripheral visual acuity will probably not be enough to recognize and localize the object. In general the cup could even be completely outside the field of view, so that eye movements would be necessary to bring the cup in the foveal region of the retina.

If the brain were able to retain all the visual information coming from the retina, then it would have a quite complete map of the surroundings, in a kind of very detailed picture-like photographic representation. Although the latter is correspondent to the daily experience, there are some practical and experimental arguments against it: first is regarding the huge amount of information would be necessary to handle even for short-time scene retention, and apart that the difficulty to dynamically update the representation over the time and as soon as change takes place into the scene.

On the other hand recent research provided strong evidence to the theory that vision is not a picture-like sampling of the scene. Such evidence comes from the "Change Blindness" phenomenon: it has been observed that in certain circumstances, humans are unable to perceive changes happening in the scene, although the changes are big and affecting to semantically reach part of the scene. The Change blindness was first associated to the saccadic suppression of perception during saccadic eye movements but it has been recently proved that the blindness to change is also occurring when eye if fixated on the point of change[101].

The most common change detection paradigm is the so called "flicker paradigm" in which two images, one representing the full scene, the other the scene missing of a meaningful part, are shortly shown on a screen alternated by a brief gray field (scene presented for 240ms and gray field for 80ms). Experiments proved that the changes were hard to be detected even if participant were asked to report the occurrence of any change to the scene, and to report it by pressing a button, as soon as the change was perceived. In some cases participants needed a continuous stimulus repeated up to 1 minute before becoming aware and able to report the change in the scene. The phenomenon has been proved to be a quite general one, not depending on the particular kind of images, not dependent to the kind of blanking interval and colors and occurring for a wide variety of changes [91, 90].

Change blindness has first be used to sustain that we do not construct any detailed picture-like internal representation of the scene [77]. Because we have highly mobile eyes we can point directly to the world to extract the information we need and the low cost of this would make not necessary to interrogate any internal representation. [76].

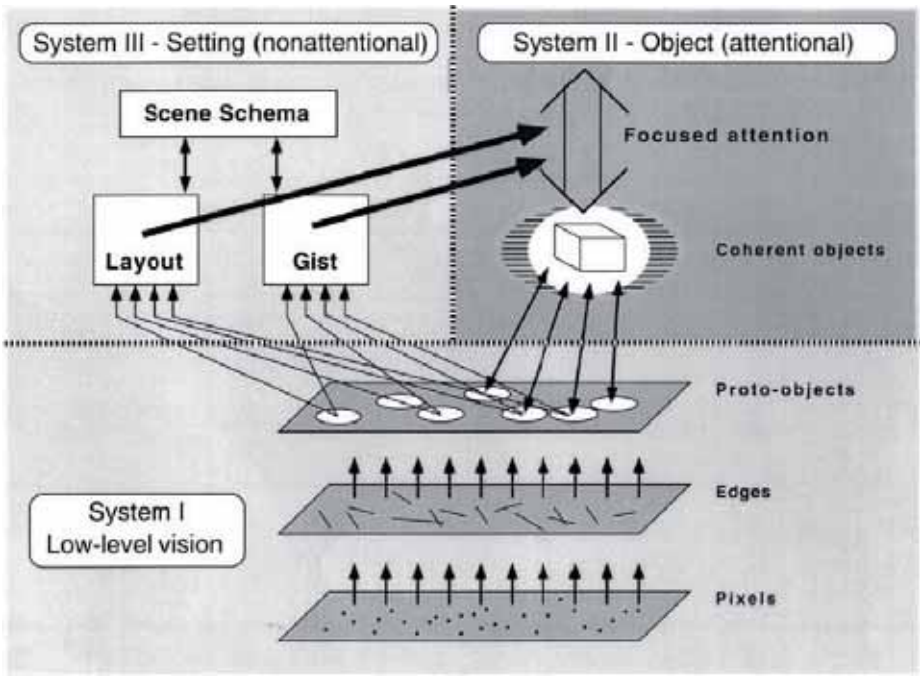



**Figure 2.4:** Triadic model of attention. From Rensink [91]

And although the motion signals associated to the change should attract attention to specific location making changes visible, the observed inability to perceive changes in the scene in presence of a brief blanking, could be explained as a result of the induced motion signal swamping the change signal. A possible solution is given by the Triadic Model of Rensink and described in the following section 2.2.1.

### 2.2.1 Triadic architecture of Rensink

Building on top of the experimental evidence coming from the flicker paradigm [91], Rensink suggests that little of the retinal information is stored in the brain, so vision is the result of a dynamic process based on a dynamic representation of the perceived world, in which humans can only attend a very limited part of the scene available at a certain moment, although the daily experience is to have detailed and complete representations of the scene.

Vision is described by Rensink as the result of the interaction of three largely independent systems. An Early-level process receives continuously new retinotopic information and continuously creates detailed and volatile proto-objects. These proto-objects, consisting of some aggregate information, constitute the only available information that is passed to the higher attentional and non-attentional systems.

The non-attentional system or "Setting system" performs an overall assessment referred as the Gist and Layout of the scene. The Gist is thought to account for the

kind of the scene, the Layout is instead related to the spatial organization of proto-objects. The Setting system is than more stable over the time than the proto-objects and is thought as the part of the Vision system responsible of the making a stable perception of the scene out of the continuous flux of proto-objects. Gist and Layout describe the whole scene at the first sight, even before any of its parts have been perceived at a high resolution, and are supposed to be quickly assessed from a low resolution input [73].

The attention system or Object system is the responsible to form a stable object representation. According to the "coherence theory" the focused attention give rise to the perception of the object by keeping together the proto-objects with high degree of coherence over time and space. As soon as the focused attention releases the proto-objects, the object loses its coherence.

The model does not provide a computational procedure to implement such focused attention mechanism, and the description is provided in figurative terms "as a hand collecting some of the proto-objects". On the other hand it is very clear that the attentional system needs to be guided by the Gist and the Layout will provide a guidance (or in different words we would say providing a context) conditioning which proto-objects to grab next.

### 2.2.2 Virtual representation

From the point of view of explaining Vision an interesting question regards "the stable perception of the world" how humans perceive the scene as complete, coherent and rich in details, although scenes are never seen completely.

This is well explained by the model of Rensink in terms of Virtual Representation that is thinking Vision as a "just in time process able to provide detailed representation of the scene whenever required". In such a way humans do not notice these limitations of their own Visual system, because of the coordination of the sub systems of the triadic model (and in particular the coordination of attention and selection of proto-objects) allows to get the right information from the world just when it is needed.

The Virtual representation is thought in analogy to the way computer networks are designed, and computer appears to contain all the information available in the network as long as the information can be instantly accessed when requested. It's a powerful scheme in which the complexity of the system is dramatically reduced at the cost of having an partial representation that might sometimes turn into an incoherent one.

## 2.3 Computational models of attention

Many different computational models of attention have been presented. In a recent survey have been listed more than 65 models, addressing visual attention as a saliency based mechanism [12]. Attention is generally thought as a combination of scene driven bottom-up factors and cognitively task related top-down factors. Here we look first to the pure saliency based models and than we look at the current attempts to model top-down constrains.

### 2.3.1 Saliency based models

The Feature Integration Theory by Treisman & Gelade [117] first described a way to account for the way human attention is attracted by features. Such theory constituted the basic building block for many attention models, first the model proposed by Koch and Ullman [60] which formalized the theory in a feed-forward model combining intensity, color, orientation features to create a *Saliency Map* representation.

The model proposed by Itti [53, 52] include a multilevel pyramidal representation of the input image and features are extracted at all the pyramid levels from the color channels (red, green, blue, yellow) and from the intensity, and local orientation maps. A center-surround mechanism computes within-map differences creating pyramidal *feature maps* that are combined and normalized to create the *conspicuity maps*. The *Saliency Map* is than the linear combination of the conspicuity maps. Several weighting factors modulate the contribution of the conspicuity and feature maps. A winner-take-all neural network selects the most salient location and an inhibition of return mechanism allows to detect all the salient locations.

The concept of saliency map proved to be useful to predict the location likely to be attended and has been largely used, providing a topographic map of locations of interest. Several models have been developed using the basic scheme by Itti. Several kinds of features have been used to give rise to conspicuity maps and different ways to combine the features have been explored. All those computational models focus on the bottom-up contribution to attention and are here referred as the *saliency based* computational models of attention.

Bruce and Tsotsos [15] proposed a bottom-up attention model based on Information Maximization, using the Shannon's self-information measure for calculating saliency as $-log(p(f))$, where $f$ is a local visual feature vector. Itti and Baldi [51] defined a Bayesian surprise map, measures how data affects an observer, in terms of differences between posterior and prior beliefs about the world. Only data observations which substantially affect the observer's beliefs are accounted as long as yielding surprise. Surprise map are irrespectively of how rare (or informative in Shannon's sense) are the data observations. Kienzle et al. [57] proposed to learn a visual saliency model directly from human eye movement data. Instead of using Gabor or Difference-of-Gaussians filters, they directly use image intensities to train linear classifier. The saliency function is than determined by the fact that it maximizes the prediction performance on the observed data. The advantage of this approach is that the features are not predefined in the system.

### 2.3.2 Top-down modelling

Top-down factors are universally accepted to be an important modulating factor of eye moments, and complex cognitive process are thought to be responsible of such effect that. Pioneering work by Yarbus 2.1.2 proved that the task at the hand and the semantic content of the scene, have a dramatic effect on the way humans look at the scene and the visual search (specifically we observe the scanpath) changes among different tasks. Some computational models have tackled the problem of modelling top-down factors by 1) modulating the model's weights 2) accounting for the context 3) accounting for objects:

**Modulating the model's weights.** A way to take into account the top-down constraints (attention) is to modulate the weights in a saliency based model to bias the allocation of attention towards the object of interest. When looking for a green vertical object a higher contribution could be assigned to greenish color features and to almost vertical directions. Such approaches take inspiration from the guided search theory [130, 131] initially developed by Wolfe. Similarly Desimone and Ducan [29] account for the top-down factors by modulating the weights of the model and making different balance of how the conspicuity maps contribute to the final saliency map. More recently the discriminant saliency approach described by Gao [37] derives an optimal saliency detector by looking at the discriminant power of a set of features in a center-surround mechanism with respect to a two-class classification problem: the stimuli of interest and the non interest stimuli. Saliency decisions are taken in a optimal decision-theoretic sense, comparing the classified locations and taking the lowest expected probability of error.

**Accounting for the context.** A different way to take into account top-down contribution is by acknowledging the modulating power of the context in how humans look at the locations of their interest. Key questions then are about how to model the top-down contributions and how they combine with the bottom-up ones.
Torralba and Oliva [115, 114, 116, 75] combine bottom-up saliency contribution and the top-down scene context. In the contextual guidance model, the saliency is derived as $P(f|F_G)$ where $F_G$ represents the global image features and is related to the probability of presence of the target object in the scene (gist). Contextual contribution instead is modelled as a learned association between target locations and global scene features. The saliency information is so modulated by the scene prior and the two maps are combined up according to a learned weight to avoid that the product could be constantly dominated by one factor. Such a model outperforms a purely saliency based model in predicting human fixation locations in a search task.
In the work by Zhang et. al. [135, 134], the saliency of a point $z$ is defined as $P(C = 1|F = f_z, L = l_z)$, where $C$ denotes the class label associated to $z$, $L$ denotes the location of a point, F be the visual features. Using the Bayes rule and making use of conditional independence hypothesis the saliency of a point comes of to be defined by three components: a self information of the features making rarer features the most informative, the likelihood term that favors feature values that are consistent with the knowledge of the target and third a location prior independent of visual features and reflects any prior knowledge of where the target is likely to appear.
Ehinger et al. [32] use a scheme similar to one used by Torralba to account for human eye movements in people search task in natural scenes. They jointly model the contribution of bottom-up saliency, gist, and object features by linearly integrating the three sources of guidance: $M(x, y) = M_S(x, y)_1^\gamma + M_T(x, y)_2^\gamma + M_C(x, y)_3^\gamma$ in which $M_S(x, y)$ is the bottom-up visual saliency, $M_T(x, y)$ the learned visual features of the target's appearance, and $M_C(x, y)$ a learned relationship between target location and scene context. The exponents $(gamma_1, gamma_2, gamma_3)$, act as weights and are required and learned from the data to avoid that the combined distribution could be dominated by one of the sources.

**Accounting for objects.** The top-down constraints can then be accounted by directly looking to the presence of specific objects. Recent research proved that faces

as well as text have strong attractive power [123], by eye movements measuring it was observed that objects predict fixations better than saliency maps built on early features such as color, contrast, orientation, motion [33]. The work by Cerf et. al. proved that text and faces attract fixation and independent of the task, are even difficult to be ignored in free viewing condition. Enhancing the model of Itti by the faces and text conspicuity maps, as coming from trained face and text detractors, improves the eye fixation location prediction [19, 20, 21].

**Other.** Some other approaches aimed to describe vision as inference on a graphical model. Chikkerur et al. [25] proposed a model of attention similar to the model of Rao et al. [86] to jointly model features, object identity and locations, in which attention emerges as the inference in a Bayesian graphical model. The model proved to explain well some spatial-attention psychological mechanism as well as predict human fixations in free viewing and search tasks.

## 2.4 Criticism to the saliency based models

Saliency based models have been used to predict eye fixations quite successfully in artificial stimuli arrays. To a large extent, the psychological literature was conceived on simple stimuli, nevertheless the key role that the above models continue to play in understanding attentive behaviour should not be overlooked as long results from psychophysics experiment do not directly extend to real scene. It has been proved in the specific case of natural images, and in particular those for which no related top-down task is involved, the features seem to be discriminative enough to reliably predict the fixation locations as proved by work on saliency modelling [53] [52].

Although when moved into more complex stimuli, like street view pictures, these models do not predict well fixation locations. The adoption of complex stimuli has sustained a new brand of computational theories, though this theoretical development is still at an early stage [36]. Human generated scanpath exhibit inter-subjects and intra-subjects variability, are dependent to the image stimulus in a saliency-based attention capture fashion, dependent on the task at hand [132] and, more in general, perception of complex scene is based on cognitive task that may completely override saliency. To this extent it's not surprising that nobody has really succeeded in predicting the sequence of fixations of a human observer looking at an arbitrary scene.

**When saliency fails to predict eye movements** In walking experiments saliency based attention models do not predict well fixations, Rothkopf and Ballard showed that in artificial walking experiments in which participants had to avoid obstacles [96], most of the fixations are direct to the objects and not to the background scene as predicted by saliency. Similarly in ball sports the expert players make saccades to the places where they expect the ball [4] as well as in sandwich making [43] some of the fixations are even directed to empty space in relation to where the object will be placed.

**Objects predict fixations.** Einhäuser, Spain and Perona [33] eye tracked human observers while observing photographs of common natural scenes and proved that fixated and control locations can be better distinguished by object-level information than by image saliency, bringing evidence of how the early saliency plays a not so

central role in fixation generation.

**Saccadic bias.** From a perception point of view it's not clear the degree to which eye movements maximize the information gain (as discussed in Sec. 2.1.1). In this respect Tatler observed that human and model generated scanpaths substantially differ in their statistical properties, specifically the saccade amplitude plots present quite different characteristics revealing how the saliency model does not capture the underlying mechanisms of human perception [110]. From a quantitative point of view Tatler showed that the saliency models are able to predict eye fixations just a little better than chance [111]. Employing an edge density classifier Tatler measured the performance as a proportion of correct classification and proved that including systematic tendencies will speed up the system performance.

Saliency based models predict well eye fixations when a large visual signal present in the environment acts as a proxy for visual attention. However when moving to the real world such large signals are no more present or masked by a number of signals acting at the same time. Tatler argues against the dominant role of saliency [110] and about the role of such large signals and how often attention is captured by these large signals in ordinary oculomotor behavior.

**Picture-viewing paradigm.** There is than a methodological problem related to the measurement of eye movements. The picture-viewing paradigm has long been used to investigate about how humans gaze in natural environments, however some bias may be introduced by the paradigm itself more than the visual stimulus. The central bias [109] observed in the fixation distribution may be due to the framing constraint introduced by the monitor used for the stimulus display. In addition the dynamic range of a picture is much less than a real scene, motion cues and many depth cues are absent, the observer viewpoint is fixed and decided by the viewpoint of the photographer introducing compositional biases. This rises the question about the use of picture-viewing paradigm and its general validity in extending results to the vision in real world.

**Inhibition of return.** A practical problem regarding the use of saliency based model in dynamical scenes has been addressed by Henderson [45]. Since the traditional model works with static images, the saliency map can be computed pre-attentively and all the successive saccades be computed on the basis of such a map. However in a dynamic scene the saliency map has to be updated or computed several times. So a single map could be retained over multiple fixations, or a new map can be computed for each successive fixation. In the latter case it rises the necessity to have a mechanism able to handle the inhibition of return across a number of fixations and to assign to the correspondent point in the newly generated map. In general the inhibition of return even in static images poses a big problem in the possibility of having re-fixations. A transient inhibition will allow cyclic repetition of the scanpath leading to unnatural repetitive scanpaths, instead a long lasting inhibition of return will make it impossible to reproduce the typical human behavior. E.g. several re-fixations of a location are observed by Ballard in a block-copying task [5].

### 2.4.1 The layered model of Schütz

In the light of a series of limits shown by the use of saliency to predict fixations, it raises the question of how and when it happens that saliency can be overridden by top-down constraints? and how these constraints should be modelled?

Schütz suggests that the control of saccadic target selection could be produced by a layered model in which separate factors act at different levels of visual processing [97], in which the saliency is only one of the factors. A better explanation of visual processing can be given in terms of several modulating mechanisms contributing to the fixation location selection at different levels. These mechanisms are related to the concept of object, value and plans.

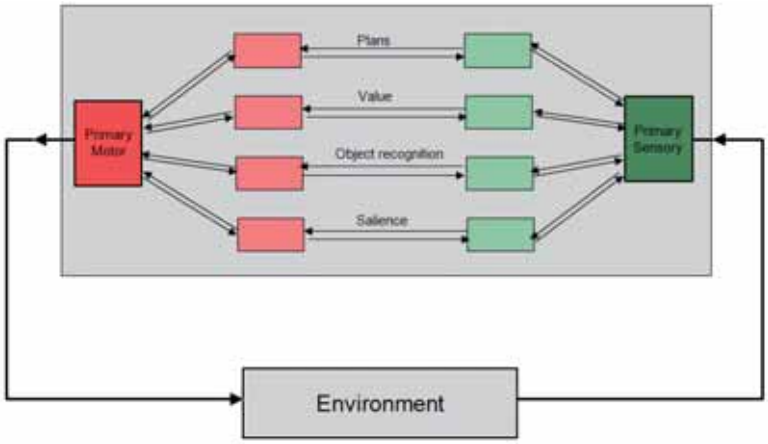



**Figure 2.5:** The model of Schütz. Attention deployment described at several levels of visual processing.

Individual maps may exist for each factor and the levels of saliency, object, value and plans might be somehow integrated through a number of local interactions in order to make a common priority map closely linked with the saccadic eye movements control. The interaction among the different levels of processing should be the a research priority to investigate perception and eye movements.

### 2.4.2 The Object level

Object could predict fixation locations better than saliency in a number of real world task. Thinking to the objects as material things would allow to move in an action-perception loop in which eye movements are basis object manipulation and scene interaction. In such a context saccadic target selection could be driven by the object's suitability for a given task. On a compartmental line of thinking the abstract concept of familiarity and suitability of an object could be claimed to justify specific eye movements, as well as more skilled or less skilled oculomotor behavior. In a real world experiment involving the interaction with objects it's quite intuitive to think of the objects itself as the triggers for saccadic target selection. The only features without

the concept of object would not allow to justify some task related eye movements.

Support to this theory comes from the work of Perona proving that object plays a relevant role in fixation generation [33] fixated and control locations can be better distinguished by object-level information than by image saliency. Under specific tasks such as artificial walking [96], text reading or people counting search task the saliency contribution is completely overridden by the contribution of the object.

It's well known that faces and text in general are playing an important role in saccade control regardless to the task [19]. Than a practical approach is to enhance saliency map by object detection algorithm to improve the fixation prediction. Cerf et. al. [20, 21] found that using face detection algorithm to extend the saliency map led to improved gaze prediction for images containing faces.

**Object priming.** Some neurophysiological studies have shown by making Electroencephalography (EEG) recordings that humans are capable of detecting very rapidly the presence of animals (or other objects) in a scene; the time for such detection is less than the time needed to make a saccade [113]. Recently work of Drewes, Trommershäuser, and Gegenfurtner [30] have shown that humans are able to make animal detection and estimated the visual processing time in 120 ms. Such detection were performed on realistic photography containing only one animal each and authors observed that observers were able to saccade to the animal directly and in many cases the saccades were directed to the animal's head. They also showed that a saliency-based algorithm such as the computational one by Walther and Koch [122] cannot account for the human fixation performance.

### 2.4.3 The Value level

Studies have been done in learning theory accounting for the consequences of hands and body movements to actively manipulate the environment with immediate positive or negative consequences, nevertheless a similar investigation is lacking on the side of eye movements. Schütz observed that a fundamental contribution to the selection of saccade location should come from the value of the fixating to a specific location. The value of an eye movement is understood as the positive or negative consequence of making a change to the environment. At this point it's worth underline that eye movements, although not directly making manipulation of the environment, they determine changes in the field of view, that is affecting to which portion of the scene to look at than which part of the scene take into account. Only the fixated part will be attended and processed, the other parts will be ignored. In this sense eye movements select the visual information and be able to make good eye movements allows to gather relevant information from the scene. Value is than related to a possible reward from the having done good eye movements.

Navalpakkam, Koch, Rangel, and Perona [72] investigated the impact of value and saliency on choice. In psychophysics experiments, human subjects attempted to maximize their monetary earnings by quickly picking items from a brief display containing distractors and two salient and valued targets. Observers picked the target that maximized the expected reward, not the more salient target nor the more valuable target, although decisions are affected by both saliency and value. Results proved to be consistent with the predictions of an ideal Bayesian observer

From this perspective the contribution of the Value assumes the characteristics of a mixed bottom-up and top-down mechanisms. According to this way of thinking the next fixation location is only partially driven by saliency. We are embracing a more dynamical view of perception, a view that is accounting for an action-perception loop, and suggesting to think of Value as a predictive mechanism able to evaluate different possible eye movements, and aiming to gather (the most) relevant visual information needed to better perform in the environment.

## 2.5 Systematic tendencies

Eye saccadic movements are not randomly distributed, different driving mechanisms have been hypothesized to be responsible of the driving of fixation location selection. Only recently some research has been directed to assess the capability of the human visual system to perform eye movement from a biological point of view. Tatler and Vincent argued that as well as certain combination of finger movements are much more frequent than others [50], in a similar way certain eye movements should be more likely than others and that such knowledge on behavioral bias in eye movements might have a high informative content by itself [111].

Eye movement recordings reveal that human generated scanpaths have some statistic regularities: saccades follow a positively skewed distribution of amplitudes and the horizontal and vertical directions seem to be preferred to diagonal saccadic directions. Tatler names those regularities as oculomotor biases, since they seems to be due to specific characteristic of the visual system.

Saliency based models (e.g. the Itti model) produce scanpaths with statistical distributions (saccades amplitudes and orientation) completely different from the human ones. Saccade amplitude looks Gaussian distributed and no long tails are observed, while saccadic direction are equally distributed in all directions [110].

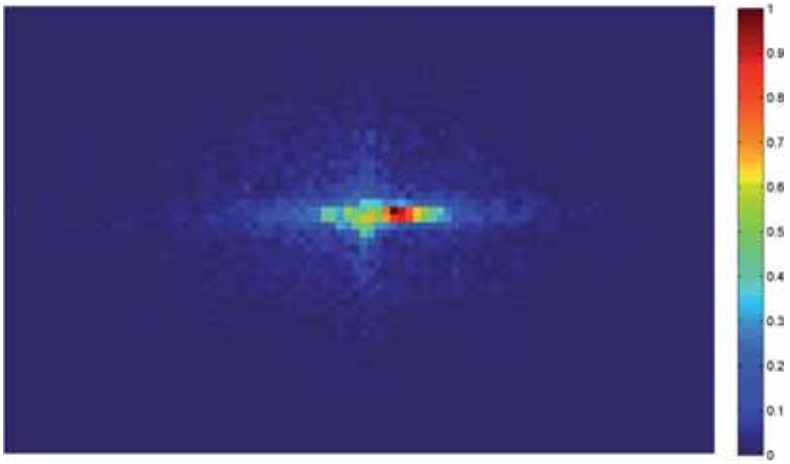



**Figure 2.6:** Amplitude distribution of gaze shifts. Plot of the data from eye tracking experiments on 6 subjects and 110 images.

These saccadic biases can be further analyzed in relation to the direction and magnitude of a saccade. Humans have a clear bias in making horizontal saccades instead of vertical ones, and to make vertical saccades more frequently than oblique saccades. Moreover there were more longer saccades in the horizontal direction than for the overall distribution. In contrast vertical saccades are more frequently of smaller amplitude than either the overall distribution or horizontal saccades. See Fig. 2.6.

These saccadic biases cannot be ignored by computational models of attention aiming to predict fixation locations, Tatler and Vincent show that a model based only upon systematic tendencies outperforms saliency-only based methods. The best performance is obtained putting together saliency and systematic tendencies.

Such systematic tendencies are responsible for positively skewed saccade length distribution as well as the non-isotropic direction distribution. The resulting oculomotor behavior allows local search on interesting locations and then wide relocation to new locations performing a human inspired optimal search. Such resulting behavior is similar to the animal foraging behavior.

## 2.6 Discussion

Let us summarize the main points tackled in this chapter. First we introduced a review on eye movements and put it in relation with the peculiarity of the limited foveal visual acuity. To this end the saccadic eye movements seems to be finalized to bring the fovea on the part of the scene we need to see clearly. However it's not clear up to which extent the eye movements follows an information maximization strategy, as supported by Geisler. Eye movements seems to follow complex strategy and the only perceptual stimulus seems to be not sufficient to explain the eye movements in the large. Pioneering investigation by Yarbus showed that recordable human eye scanpath are largely affected by the task. Different scanning pattern are observed in relation to the information needed to be extracted, asking to account for more general concept as the scene context and the relationships between objects.

Further investigation in the psychological field addressed the problem of internal representation of the scene. From such point of view it seems that human do not retain a picture-like representation as it would be incompatible to the observed Change blindness phenomenon. In such sense vision is more likely to maintain an intermediate representation to which Rensink refers as proto-objects. In this respect we take the Triadic Model of attention as a general scheme towards the implementation of computational model of attention able to account for eye movements in complex and cluttered scenes.

Current saliency based models address well artificial stimulus array and proved good performance in natural scene with little or no context. However it emerges the need of models able to better predict fixations in cluttered environments such as street view images. How the top-down constrains should be modelled and how they could interact with stimulus driven saliency are open problems under debate. Some progress in this direction have been made by approaches aiming to embed object detection in the saliency description, and the approach of Torralba accounting for the Gist as a top-down mechanism able to modulate the bottom-up saliency-map description.

From the cognitive science field some progress in modeling of top down constraints have been recently made by Schütz calling for a layered model of attention in which Saliency is treated as one of the factors guiding attention. Visual attention could be than explained by the interaction of several modulating mechanism, such as Objects, Value, Plans and Saliency. This is in agreement with several studies which led evidence to the role of objects in attracting attention, although from the computational point of view it remains open the problem of how to make a rough object detection from low resolution images. The Value is, instead, a seldom used concept in computer vision, although promising as allowing to close the action-perception loop. In such a sense the Value can be interpreted as a predictive mechanism able to evaluate different possible eye movements, and aiming to select the best one.

Complementary to this some recent investigation showed the existence of some systematic tendencies in eye movements. Tatler et al. proved that only accounting systematic tendencies is eye movements would allow to predict eye movements better than using the only saliency. Novel models of attention should be able to include such oculomotor bias and use them as efficient/human inspired search mechanism. In this sense it seems to delineate two components: a perceptual one, as described at multiple levels of Objects, Value, Plans and Saliency, and a motor one described in terms of eye movements systematic tendencies. We propose that the model of Rensink might be the glue between the perceptual and the motor component, and such modeling will allow to better describe the mechanism behind the attention allocation in street view images.

# Chapter 3

# Modeling Task-dependent Eye guidance

In this chapter, we introduce a model of attentional eye guidance based on the rationale that oculomotor control depends on two interacting but distinct processes: an attentional process that assigns value to the sources of information and motor process that flexibly links information with action. In such framework, the choice of where to look next is task-dependent and oriented to classes of objects embedded within pictures of complex scenes. The dependence on task is taken into account by exploiting the value and the reward of gazing at certain image patches or proto-objects that provide a sparse representation of the scene objects. The different levels of the action-perception loop are represented in probabilistic form and eventually give rise to a stochastic process that generates the gaze sequence. This way the model also accounts for statistical properties of gaze shifts such as individual scanpath variability.

## 3.1 The model

In the light of the discussion provided in Chapter 2, it is convenient to phrase the *Where to look next?* question in the language of stochastic processes. To such end, we represent the sequence of gaze positions through the time-varying random variable (RV) $\mathbf{r}_{FOA}(\cdot)$, and the problem turns into the issue of how to sample the new gaze position $\mathbf{r}_{FOA}(t+1)$ when at time $t$ gaze is deployed at $\mathbf{r}_{FOA}(t)$, the latter being the center of the focus of (overt) visual attention (FOA). In other terms, the transition $\mathbf{r}_{FOA}(t) \rightarrow \mathbf{r}_{FOA}(t+1)$ is a transition whose dynamics is that of a stochastic process.

In this perspective, denote $\mathcal{A}(t)$ the ensemble of time-varying RVs defining the oculomotor action setting, while $\mathcal{W}(t)$ stands for the ensemble of time-varying RVs characterising the scene as actively perceived by the observer. We are interested in knowing the probability of shifting the gaze to the new location $\mathbf{r}_{FOA}(t+1)$, namely $P(\mathbf{r}_{FOA}(t+1)|\mathcal{A}(t), \mathcal{W}(t), \mathbf{r}_{FOA}(t))$ based upon all the information that the visual system has available to it, that is the current gaze location $\mathbf{r}_{FOA}(t)$, the scene $\mathcal{W}(t)$ as perceived from image $\mathbf{I}$ gazed at $\mathbf{r}_{FOA}(t)$, the oculomotor action setting $\mathcal{A}(t)$ chosen under the given task $\mathbf{T}$.

To solve this problem, our model relies on the following assumptions:

- The scene that will be perceived at time $t+1$, namely $\mathcal{W}(t+1)$ is inferred from the raw data, here in the form of a picture $\mathbf{I}$, gazed at $\mathbf{r}_{FOA}(t+1)$ under the task $\mathbf{T}$ assigned to the observer, and is conditionally dependent on current perception $\mathcal{W}(t)$; thus, the perceptual inference problem is summarised by the conditional distribution $P(\mathcal{W}(t+1)|\mathcal{W}(t), \mathbf{r}_{FOA}(t+1), \mathbf{I}, \mathbf{T})$;
- Task $\mathbf{T}$ being assigned, the oculomotor action setting at time $t+1$, $\mathcal{A}(t+1)$, is drawn conditionally on current action setting $\mathcal{A}(t)$ and the perceived scene $\mathcal{W}(t+1)$ under gaze position $\mathbf{r}_{FOA}(t+1)$; thus, its evolution in time is inferred according to the conditional distribution $P(\mathcal{A}(t+1)|\mathcal{A}(t), \mathcal{W}(t+1), \mathbf{r}_{FOA}(t+1), \mathbf{T})$.
- The action setting dynamics $\mathcal{A}(t) \rightarrow \mathcal{A}(t+1)$ and the scene perception dynamics $\mathcal{W}(t) \rightarrow \mathcal{W}(t+1)$ are intertwined with one another by means of the gaze shift process $\mathbf{r}_{FOA}(t) \rightarrow \mathbf{r}_{FOA}(t+1)$: on the one hand next gaze position $\mathbf{r}_{FOA}(t+1)$ is used to define a distribution on $\mathcal{W}(t+1)$ and $\mathcal{A}(t+1)$; meanwhile, the probability distribution of $\mathbf{r}_{FOA}(t+1)$ is conditioned on current gaze position, $\mathcal{W}(t)$ and $\mathcal{A}(t)$, namely $P(\mathbf{r}_{FOA}(t+1)|\mathcal{A}(t), \mathcal{W}(t), \mathbf{r}_{FOA}(t))$.

By fulfilling such assumptions, the actual shift can be summarised as the statistical decision of selecting a particular gaze location $\mathbf{r}^{\star}_{FOA}(t+1)$ on the basis of $P(\mathbf{r}_{FOA}(t+1)|\mathcal{A}(t), \mathcal{W}(t), \mathbf{r}_{FOA}(t))$ so to maximize the expected payoff with respect to the given task $\mathbf{T}$.

The conditional dependencies between RVs $\mathcal{A}(t), \mathcal{A}(t+1), \mathcal{W}(t), \mathcal{W}(t+1), \mathbf{r}_{FOA}(t), \mathbf{r}_{FOA}(t+1), \mathbf{T}, \mathbf{I}$ can be explicitly represented by means of the Probabilistic Graphical Model (PGM) depicted in Fig. 3.1. A PGM [61] is a graph-based representation where nodes denote RVs and arrows code conditional dependencies between RVs. It is important to note that arrows do not generally represent causal relations, though in specific situations it could be the case. More precisely, the structural dependency $X \rightarrow Y$, states the probabilistic dependency of RV $Y$ on $X$ represented via the conditional probability $P(Y|X)$). Indeed, this is one suitable way of formalising a model at the computational theory level [59].

Note that the scheme in Fig. 3.1 can be read as a dynamic (time-varying) PGM [61]. Further, it is important to note that the state transition dynamics of the RVs from time $t$ to time $t+1$ only depends on the state of such RVs a time $t$. In the language of stochastic processes this statement characterises a first order Markov process. Such formal assumption, which is largely exploited in dynamic PGMs [61] is occasionally summarised as a memoryless assumption about the process. By analogy with the psychological literature, this would amount to say that our model when used to perform a search task, implements a kind of visual search that has no memory [48]. However, such liberal interpretation turns to be improper. A first order Markov assumption about the gaze shift $\mathbf{r}_{FOA}(t) \rightarrow \mathbf{r}_{FOA}(t+1)$ only states that the transition probability has the following property: $P(\mathbf{r}_{FOA}(t+1)|\mathbf{r}_{FOA}(t)) = P(\mathbf{r}_{FOA}(t+1)|\mathbf{r}_{FOA}(t), \mathbf{r}_{FOA}(t-1), \mathbf{r}_{FOA}(t-2), \cdots)$, namely, at time $t$ the probability of the transition $\mathbf{r}_{FOA}(t) \rightarrow \mathbf{r}_{FOA}(t+1)$ can be computed by conditioning on $\mathbf{r}_{FOA}(t)$, and earlier terms - at times $t-1, t-2, \cdots$ -

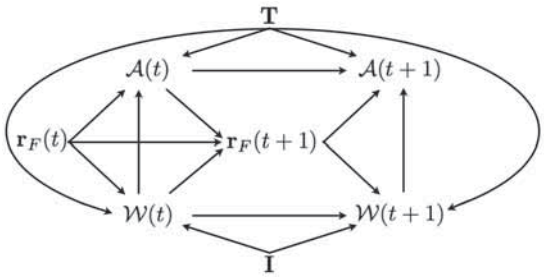



**Figure 3.1:** The model represented as a dynamic Probabilistic Graphical Model. $\mathcal{A}(t)$ stands for the ensemble of time-varying random variables (RVs) defining the oculomotor action setting (for short, the *action* component); $\mathcal{W}(t)$ is the ensemble of time-varying RVs characterising the scene as actively perceived by the observer (the *perception* component). The gaze shift $\mathbf{r}_{FOA}(t) \rightarrow \mathbf{r}_{FOA}(t+1)$ ties the dynamics of both components, and the scan path $\{\mathbf{r}_{FOA}(1), \mathbf{r}_{FOA}(2), \cdots\}$ is the result of an action-perception loop performed by the observer on an input image $\mathbf{I}$ under a given task $\mathbf{T}$. Here, the evolving loop is unrolled for two time slices, respectively, $t$ and $t+1$.

need not be taken into account. The same holds for $P(\mathcal{W}(t+1)|\mathcal{W}(t), \mathbf{r}_{FOA}(t+1), \mathbf{I}, \mathbf{T})$ and $P(\mathcal{A}(t+1)|\mathcal{A}(t), \mathcal{W}(t+1), \mathbf{r}_{FOA}(t+1), \mathbf{T})$. However, as we will discuss later, there are RVs in the sets $\mathcal{W}(t), \mathcal{A}(t)$ that are used to define probability distributions over the image spatial support (for example, the priority map and the value map represented through the spatially defined RVs $\mathbf{L}(t)$ and $\mathbf{V}(t)$, respectively) that, though evolving in time according to a first order Markov dynamics, keep track of events previously occurred. Thus, when engaged in a search task the gaze sampling mechanism may behave very differently from a sampling with no memory (i.e., with replacement [81]).

We consider two tasks: a general "free-view" task ($\mathbf{T} = FV$) and a "look for x" ($\mathbf{T} = S$) or search task. Hence $\mathbf{T}$ is a binary RV influencing, at any time $t$, both the perceptual ensemble $\mathcal{W}(t)$ and the action ensemble $\mathcal{A}(t)$. This will be obtained at the perceptual level by conditioning on task the prior probability of gazing at certain objects within the scene, while at the action level, the task will modulate the probabilities of the value and the payoff related to a possible oculomotor act. In the following sections, we will provide concrete examples of the top-down role played by the task variable $\mathbf{T}$. Further, we instantiate and discuss the actual RVs characterising the general representational levels that we have summarised through the ensembles $\mathcal{W}(t)$ and $\mathcal{A}(t)$, together with their dependencies. As a result, the PGM presented in Fig. 3.1 will be eventually specified in a full probabilistic model that we introduce in Fig.3.2 below.

For explanatory convenience, we will start our discussion from the representations underpinning the perceived scene $\mathcal{W}(t)$, as available by "freezing" the loop at time $t$ (Fig. 3.2) when gaze is deployed at $\mathbf{r}_{FOA}(t)$. Nevertheless, it is important to note that in this article we are not committing to any specific visual procedure, inasmuch as it serves the purpose of supporting the computational theory of the integrated loop.

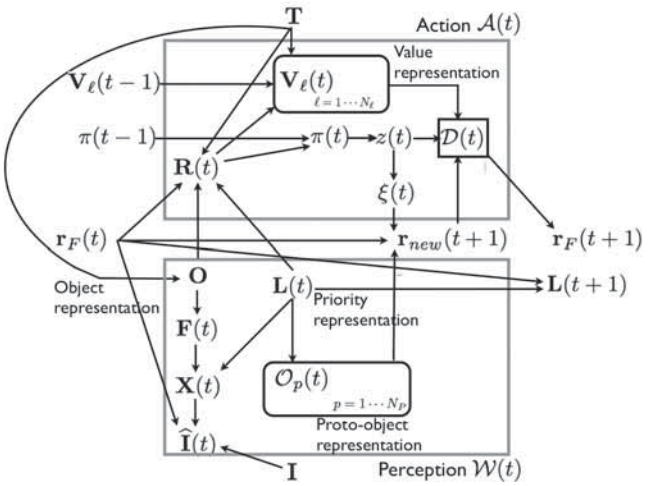



**Figure 3.2:** A snapshot of the model when gaze is deployed at $\mathbf{r}_{FOA}(t)$. It provides a detailed view of the time slice $t$ outlined in Fig. 3.1. Rounded boxes are "plates" denoting stacks of multiple random variables generated from the same distribution.

### 3.1.1 Moment-to-moment scene perception $\mathcal{W}(t)$

Consider the PGM specification of the model outlined in Fig. 3.2 and in particular the perception component at the bottom of the scheme. At time $t$, the perceived scene $\mathcal{W}(t)$ is an ensemble of different representations, namely

- $\{\widehat{\mathbf{I}}(t), \mathbf{X}(t)\}$: the *visual front-end* given by the foveated image $\widehat{\mathbf{I}}$ and a local feature map $\mathbf{X}(t)$ [25];
- $\mathbf{L}(t)$: a *priority map*, that is a map of visual space constructed from a combination of properties of the external stimuli, intrinsic expectations, contextual knowledge [25, 114];
- $\mathcal{O}(t)$: an ensemble of *proto-objects* or patches [10, 127, 122];
- $\{\mathbf{O}, \mathbf{FOA}(t)\}$: an *object-level* representation, as determined by the classes of objects that can be embedded within the scene together with the visual features characterising the appearance of such objects [25]. In this study, we take into account faces and text regions that are known to attract attention even in a free viewing task [19, 124], thus the RV accounting for objects is a binary one, *i.e.*, $\mathbf{O} = \{face, text\}$.

All together, such RVs define the joint probability of perceiving $\mathcal{W}(t)$, the task $\mathbf{T}$ being assigned, when $\mathbf{I}$ is observed after the gaze shift $\mathbf{r}_{FOA}(t-1) \rightarrow \mathbf{r}_{FOA}(t)$:

$$P(\mathbf{O}, \mathbf{FOA}(t), \mathbf{L}(t), \mathbf{L}(t-1), \mathcal{O}(t), \mathbf{X}(t), \widehat{\mathbf{I}}(t) | \mathbf{I}, \mathbf{T}, \mathbf{r}_{FOA}(t), \mathbf{r}_{FOA}(t-1)).$$

The "foraging eye", by gazing at $\mathbf{r}_{FOA}(t)$, allows the observer to gauge, at time $t$, the actual scene represented here by the given image $\mathbf{I}$ and thus to construct $\mathcal{W}(t)$. The

first step for inferring the perceived scene $\mathcal{W}(t)$ is to derive a foveated representation of the input image $\mathbf{I}$. Many visual attention models do not take into account the retinal position of image information, and decreasing retinal acuity in the periphery is surprisingly overlooked [110]. Yet, retinal anisotropies in sampling play a role in tendencies to move the eyes in particular ways, and Tatler *et al.* [110] raised the point that the assumption of uniform spatial sampling can lead to distributions of saccade amplitudes that do not match human eye behaviour. Thus, in our model the starting point is represented by the foveated image $\widehat{\mathbf{I}}(t)$, that is $\mathbf{I}$ gazed at $\mathbf{r}_{FOA}(t)$. The foveated image $\widehat{\mathbf{I}}(t)$ is structured as a pair $\widehat{\mathbf{I}}(t) = \{\widehat{\mathbf{I}}_{LR}(t), \widehat{\mathbf{I}}_{HR}(t)\}$, respectively a low-resolution (LR) one, mainly exploited during long relocations of gaze, and a high resolution one (HR), mainly used to support local fixational movements and small saccades.

From the foveated image, perception is accomplished according to a hierarchical scheme (cfr., Fig. 3.2). The structural dependencies shaping such hierarchy can be formalised in terms of probabilistic conditional dependencies among the RVs introduced above, which amounts to the following factorisation of the joint pdf introduced above:

$$\begin{aligned} &P(\mathbf{O}, \mathbf{FOA}(t), \mathbf{L}(t), \mathbf{L}(t-1), \mathcal{O}(t), \mathbf{X}(t), \widehat{\mathbf{I}}(t) | \mathbf{I}, \mathbf{T}, \mathbf{r}_{FOA}(t), \mathbf{r}_{FOA}(t-1)) = \\ &P(\mathbf{O}|\mathbf{T}) P(\mathbf{L}(t)|\mathbf{L}(t-1), \mathbf{r}_{FOA}(t-1)) P(\mathcal{O}(t)|\mathbf{L}(t)) \\ &\cdot P(\mathbf{FOA}(t)|\mathbf{O}) \cdot P(\mathbf{X}(t)|\mathbf{L}(t), \mathbf{FOA}(t)) \\ &\cdot P(\widehat{\mathbf{I}}(t)|\mathbf{r}_{FOA}(t), \mathbf{X}(t), \mathbf{I}) \end{aligned} \tag{3.1}$$

The factorization specified in Eq. 3.1 makes explicit the distributions at the different levels of representation from top to bottom: the object and object feature level, $P(\mathbf{O}|\mathbf{T})$ and $P(\mathbf{FOA}(t)|\mathbf{O})$, respectively; the priority map level, $P(\mathbf{L}(t)|\mathbf{L}(t-1), \mathbf{r}_{FOA}(t-1))$; the proto-object level, $P(\mathcal{O}(t)|\mathbf{L}(t))$; the local feature level that ties object features to prioritized locations, $P(\mathbf{X}(t)|\mathbf{L}(t), \mathbf{FOA}(t))$; the foveated image level $P(\widehat{\mathbf{I}}(t)|\mathbf{r}_{FOA}(t), \mathbf{X}(t), \mathbf{I})$.

Clearly, the probability of dealing with certain classes of objects, $P(\mathbf{O}|\mathbf{T})$ depends on the kind of images taken into account according to the task. The likelihood of spatially independent object-based features, i.e., $P(\mathbf{FOA}(t)|\mathbf{O})$, can be learned offline with any suitable technique. Indeed, it is important to note that any perceptual inference model capable of top-down, object-based analysis and representation, can serve as a suitable one for the framework presented here, provided that a priority map $\mathbf{L}(t)$ is computed. One suitable procedure could be the one discussed by Chikkerur *et al.* [25], though in the work presented here there is a conceptual difference with respect to [25] in that we consider the generation of a sequence of gaze locations. Hence, the actual input to the visual inference process is in terms of a sequence of foveated images $\widehat{\mathbf{I}}(t)$. So, for instance the inference of the priority map becomes time and gaze dependent, i.e., $P(\mathbf{L}(t)|\widehat{\mathbf{I}}(t))$ rather than simply $P(\mathbf{L}|\mathbf{I})$.

The priority level representation can be inferred from the posterior $P(\mathbf{L}(t)|\widehat{\mathbf{I}}(t))$. Note that if the features $\mathbf{X}(t)$ are observed, then $\mathbf{L}(t)$ and $\mathbf{O}$ are conditionally dependent, and prioritization is biased by objects present in the scene. Note that, in the absence of object information, $P(\mathbf{FOA}(t)|\mathbf{O}) = P(\mathbf{FOA}(t))$ and $\mathbf{L}(t)$ boils down to a classic saliency map. The attentional priority is related to both the object's saliency

and any top-down biases that influence the relative importance of that object to the subject, including the suppression of objects that have already been examined during visual search, thus playing a role in participating to the elusive Inhibition of Return (IOR) mechanism [126]. The reduction in the response to a stimulus that has been fixated essentially acts as a form of short term memory that lets the priority map keep track of which potential targets have been examined. This effect is here taken into account by letting the current $\mathbf{L}(t)$ to depend on gaze location and priority at time $t-1$, $P(\mathbf{L}(t)|\mathbf{L}(t-1), \mathbf{r}_{FOA}(t-1))$ (Fig. 3.2). This modelling choice is consistent with the finding that LIP neurons receive feedback about the selected action.

Note that the distribution on $\mathbf{L}$, $P(\mathbf{L}(t)|\mathbf{L}(t-1), \mathbf{r}_{FOA}(t-1))$, serves as a spatial prior to locate object features **FOA** on the early feature map $\mathbf{X}$. But, more generally, the priority map could also be used to take into account contextual spatial modulation of visual attention [114]. We do not consider here this problem, but integrating contextual issues in our scheme is readily done (say, in the form $P(\mathbf{L}(t)|\mathbf{L}(t-1), \mathbf{r}_{FOA}(t-1), Gist)$), and it has been experimented for a text localisation task in urban street pictures using an earlier and simplified version of the model presented here [27].

The time varying priority map $\mathbf{L}(t)$ is fundamental to organise a dynamic representation of the scene in terms of *proto-objects* [90, 127, 122, 49], which serves as the actual dynamic support for gaze orienting. They are conceived as the dynamic interface between high-level and low-level processing, a "quick and dirty" interpretation of the scene [90] . There are several possibilities to compute a proto-object representation. One way is in compact form, from either a simple [122, 49] or a more complex mid-level segmentation process (e.g., [127, 7]); an alternative is to use a sparse representation [10]. This latter option, which we embrace, will be discussed in detail in Sec. 3.2.

### 3.1.2 Oculomotor action setting $\mathcal{A}(t)$

Consider now the action component at the top of the PGM in Fig. 3.2. The oculomotor action setting $\mathcal{A}(t)$ under task $\mathbf{T}$ can be defined through the following ensemble of RVs:

- $\{\mathbf{V}(t), \mathbf{R}(t)\}$: $\mathbf{V}(t)$ is a spatially defined RV used to provide a suitable probabilistic representation of value; $\mathbf{R}(t)$ is a binary RV defining whether or not a payoff (either positive or negative) is returned;

- $\{\pi(t), z(t), \xi(t)\}$: an *oculomotor state representation* as defined via the binary RV $z(t)$, occurring with probability $\pi(t)$, and determining the choice of motor parameters $\xi(t)$ guiding the actual gaze relocation;

- $\mathcal{D}(t)$: a set of state-dependent statistical decision rules to be applied on a set of candidate new gaze locations $\mathbf{r}_{new}(t+1)$ distributed according to the posterior distribution on $\mathbf{r}_{FOA}(t+1)$.

These RVs provide different levels of representation suitable to support a value-based competition among different regions of the perceived scene serving the purpose

of statistically sampling the next gaze location. Briefly, the given task selects the most appropriate values for relocating gaze in a certain region of the currently perceived visual landscape and the possible payoffs gained after shifting. Here the landscape is summarised in terms of proto-objects. The current gaze location $\mathbf{r}_F OA(t)$ determines the actual payoff gained by the foraging eye, as a function of the availability of valuable objects at that location, which in turn is assessed through perceptual information inferred at the foveated region. The probability distribution of value defined on $\mathbf{V}(t)$ is consequently updated, while the experienced payoff biases the forager's statistical choice: to engage in local feeding or to fly away (represented through the binary RV $z(t)$). Such "coin toss" is fuelled by the competition between the time spent in local exploration and the payoff gained, which shapes the "coin fairness" parameter $\pi(t)$. At each moment $t$ a set of reachable new gaze locations $\mathbf{r}_{new}(t+1)$ is sampled so to account for both the current visual landscape, represented in terms of proto-objects $\mathcal{O}(t)$ and the motor parameters (shift angles and amplitudes as determined by $\xi(t)$) that are most plausible given the state $z(t)$. Then, as a function of current oculomotor state (feed / fly), the next gaze location $\mathbf{r}_F OA(t+1)$ is statistically selected within the set of candidate locations ranked in terms of expected payoff, thus taking the value of such locations into account. Eventually, the gaze shift $\mathbf{r}_F OA(t) \rightarrow \mathbf{r}_F OA(t+1)$ is actually performed.

#### 3.1.2.1 Value and payoff

Following the discussion in Chapter 2, we use the payoff (or reward) as an operational concept for describing the value that the foraging eye gains, under a given task, for landing in $\mathbf{r}_F OA(t)$. In an object-based setting it amounts to ascribing a value to one or more objects that can be sensed in the FOA region centered in $\mathbf{r}_F OA(t)$.

In a more formal way, we cast $\mathbf{R}(t)$ as a binary variable, with discrete values of one and zero and we assume that the probability of the *experienced payoff* $\mathbf{R}(t)$, at location $\mathbf{r}(t)$ is described by $P(\mathbf{R}(t)|\mathbf{r}(t), \mathbf{L}(t), \mathbf{O}, \mathbf{T})$. In the vein of [103], payoff magnitude is encoded as the probability $P(\mathbf{R}(t) = 1|\mathbf{r}_F OA(t), \mathbf{L}(t), \mathbf{O}, \mathbf{T})$, for which we use the shorthand $P(\mathbf{R}(t))$. Under this encoding, a gaze location $\mathbf{r}_F OA(t)$ associated with large positive payoff would give $P(\mathbf{R}(t) = 1) \simeq 1$. If the state were associated with large negative payoff , $P(\mathbf{R}(t) = 1)$ would fall near zero.

This entails that, if for generality we are going to adopt either positive or negative numerical values for payoff, we need a proper normalisation within the $[0, 1]$ interval to treat such values as probability values. Thus, following [103],

$$P(\mathbf{R}(t)) = 0.5 \left( \frac{R(\mathbf{r}_F OA(t))}{R_{max}} + 1 \right), \tag{3.2}$$

where $R_{max} = \max |R|$ is the maximal effective reward.

To compute such probabilities, the *effective payoff*, that is the actual numerical payoff assigned when gazing at $\mathbf{r}_F OA(t)$, is always computed along the feed stage and as such it is a local payoff [64]: a functional of the probability measure that is positively defined in a region centred on $\mathbf{r}_F OA(t)$ (e.g., the FOA). Clearly the effective payoff depends on the task $\mathbf{T}$. For instance, in a free viewing task, an implicit reward will be gained by observers that gaze on text or faces, due to their intrinsic attractiveness

[19, 124]. However in a "look for text" task, a higher payoff will be gained when a text region is recognised within or near the FOA centred on $\mathbf{r}_F OA(t)$.

In a more formal way let's consider $\mathbf{T}$ as a selector variable [61] that controls the multiplexed conditional probability density $P(\mathbf{R}(t)|\mathbf{r}(t), \mathbf{L}(t), \mathbf{O}, \mathbf{T})$:

$$P(\mathbf{R}(t)|\mathbf{r}_F OA(t), \mathbf{L}(t), \mathbf{O}, \mathbf{T} = S) = P(\mathbf{R}(t)|\mathbf{r}(t), \mathbf{O}); \tag{3.3}$$

$$P(\mathbf{R}(t)|\mathbf{r}_F OA(t), \mathbf{L}(t), \mathbf{O}, \mathbf{T} = FV) = P(\mathbf{R}(t)|\mathbf{r}(t), \mathbf{L}(t)). \tag{3.4}$$

Eq. 3.3 is selected when the task is a search task: in this case the effective payoff $R(\mathbf{r}_F OA(t))$ is a functional of the probability $P(\mathbf{r}_F OA(t), \mathbf{O})$ of "hitting" an object of class $\mathbf{O}$ while gazing at $\mathbf{r}_F OA(t)$. Namely, $R(\mathbf{r}_F OA(t)) = \int_{\mathcal{N}(\mathbf{r}_F OA(t))} P(\mathbf{r}_F OA(t), \mathbf{O}) d\mathcal{N}$, where $\mathcal{N}(\mathbf{r}_F OA(t))$ is a suitable neighborhood centered on current gaze location. This is basically the effective payoff locally computed in terms of a high-resolution object detector. By contrast, in a free-viewing task we compute $R(\mathbf{r}_F OA(t))$ by taking into account the local landscape of the priority map (Eq. 3.4). The rationale behind this choice stems from the fact that, although it is clear whether a subject fixates a particular region in a scene, it is not so easy to infer which features are being processed (the difference between looking and seeing [96]). In a search session fixational eye movements are likely to serve the purpose of confirming the identity of a detected object or disambiguating parts of an object; thus, the local use of a classifier/detector working at high-resolution, which is more performant than a weak and lower resolution localiser as applied in the pre-attentive stage, is a desirable choice [133]. On the other hand, the free-view task is unfortunately very uncontrolled. However, some of the highest correlations between saliency/relevance and fixation are found in free-viewing tasks. This is likely to happen, since in the absence of a specific target, visual saliency coincides with places that are useful for interpreting or remembering the scene [35]. In this case, the choice of computing the local reward as $R\left(P(\mathbf{r}_F OA(t), \mathbf{L}(t)\right)$ is a reasonable approach.

The payoff gained at $\mathbf{r}_F OA(t)$ allows to update the probability distribution of value defined on $\mathbf{V}(t)$, the time-varying spatial map of behaviourally relevant locations over the visual space, so that at each point a task-dependent value is attached. For the specific purposes of this study, we assume a layered representation of value maps, $\{\mathbf{V}_\ell(t)\}_{\ell=1}^{|\mathbf{O}|}$, in particular one for each class of objects that may be relevant for the given task. This is an extension of the scheme proposed by Navalpakkam *et al.* [72], though their study was limited to the use of primary rewards. Each location of $\mathbf{V}_\ell(t)$ represents a binary random variable $\mathbf{v}_\ell(\mathbf{r}, t)$, denoting whether $\mathbf{r}$ is a valuable point ($\mathbf{v}_\ell = 1$) or not ($\mathbf{v}_\ell = 0$).

The $\ell$-th value map at time $t^{'} > 0$ and at location $\mathbf{r}$, given the locally experienced payoff is computed as the cumulated payoff averaged on the neighborhood $\mathcal{N}(\mathbf{r})$:

$$P(\mathbf{v}_\ell(\mathbf{r}, t^{'})|\mathbf{R}(t^{'})) = k_V \left( \sum_{t=1}^{t^{'}} E_{P(\mathbf{R})}[\mathbf{R}(t)|\mathcal{N}(\mathbf{r})] + P(\mathbf{v}_\ell(\mathbf{r}, 0)) \right), \tag{3.5}$$

where $k_V$ is a suitable normalizing constant. Eq. 3.5 provides an iterative formulation of the recursive computation of the pdf $P(\mathbf{v}_\ell(\mathbf{r}, t)|\mathbf{R}(t), \mathbf{v}_\ell(\mathbf{r}, t-1), \mathbf{T})$.

At time $t = 0$, the $\ell$-th density $P(\mathbf{v}_\ell(\mathbf{r}, 0))$ is initialized as a function of $P(\mathbf{L}(t), \mathbf{O} = o|\mathbf{T})$, the object-based map obtained through a pre-attentive rough classification stage (see Sec. 3.2). The effective value at each point is assigned using Eq. 3.2. Notice that the value map is different than the priority map $\mathbf{L}$ although at $t = 0$ it might be similar since the distribution $P(\mathbf{L}|\mathbf{I})$ captures the presence of objects (in the sense of shaping an object-based top-down saliency map). Indeed, value depends on task and is adapted in time as a function of payoff: for instance in a control task, regions that are likely to contain objects do not loose value in time by always assigning positive rewards, so to be re-fixated; in a "quickly search for all objects", value of the detected object will decrease in time, since reward will be high for the first fixation on the objects and negative for subsequent fixations.

#### 3.1.2.2 Oculomotor state representation

Once a value setting is supplied, the ultimate problem of gaze relocation is to choose between feeding on local information (*intensive stage* performed through fixational movements) or "flying away" in search of more valuable foraging patches by relocating gaze (*extensive stage* via medium and large saccades) [11]. Notice that we equate fixations with local feeding, since a fixation is not simply the maintenance of the visual gaze on a single location but rather a slow oscillation of the eye (minimum 50 milliseconds duration) within a circumscribed region (typically $0.5° - 2.0°$ degrees of visual angle), [47]; longer displacements stand for saccades.

Formally we index such two states using the binary RV $z(t)$, where $z(t) = 1$ denotes the "feed" state and $z(t) = 0$ the "fly" state. We assume that after a flight (a saccade) the foraging eye is always prompted to engage in the intensive stage, that is, the transition $z = 0 \rightarrow z = 1$ occurs with probability 1. This in principle does not imply that local feeding be always actually performed: if conditions for feeding are not met and/or because of the randomness of the process, the transition $z = 1 \rightarrow z = 0$ may occur before such stage actually take place. Let $\pi(t)$ be the probability of remaining in the feeding state, $P(z(t) = 1) = \pi(t)$. Clearly, the transition $z = 1 \rightarrow z = 0$ occurs with probability $P(z(t) = 0) = 1 - \pi(t)$. In other terms, in state $z(t) = 1$ the choice of state, keep feeding or engage in a flight, can be conceived as a "coin toss" governed by the Bernoulli distribution, $Bern(z(t); \pi(t)) = \pi(t)^{z(t)}(1 - \pi(t))^{1-z(t)}$ for $z(t) \in \{0, 1\}$. The bias of such "coin tossing" procedure is, differently from [11], dependent on payoff.

The bias accounts for the competition between the time already spent within the foraging patch and the willingness of the forager to continue with local feeding. Thus, the local feeding time is evaluated through the number of points locally visited at time $t$, say $n_s(t)$; the willingness to stay or to leave is accounted for by the mean feeding rate of the forager, $\mu$, which in turn is a function of the actual payoff $R(\mathbf{r})$ gained while engaged in the intensive stage. On this basis, we model $\pi(t)$ with the exponential function,

$$\pi(t) \propto \exp\left(-\frac{n_s(t)}{\mu(R(\mathbf{r}_F OA(t)))}\right); \tag{3.6}$$

To sum, the mean feeding rate, determining the willingness of the forager to continue the feeding stage, is a function of gained payoff, which in turn depends on the

given task $\mathbf{T}$. When "biased" parameters $\pi(t)$ have been computed, the oculomotor state can be sampled as:

$$z(t) \sim Bern(z(t); \pi(t)). \tag{3.7}$$

#### 3.1.2.3 Deciding the gaze shift

The decision $\mathcal{D}(t)$ of shifting the gaze to a new position is taken in order to maximize the *expected reward* of moving to a valuable site. In our framework, the candidate new gaze locations $\mathbf{r}_{new}(t+1)$ can be obtained by sampling from the distribution $P(\mathbf{r}_{FOA}(t+1)|\mathcal{A}(t), \mathcal{W}(t), \mathbf{r}_{FOA}(t))$:

$$\mathbf{r}_{new}(t+1) \sim P(\mathbf{r}_{FOA}(t+1)|\mathcal{A}(t), \mathcal{W}(t), \mathbf{r}_{FOA}(t)) \tag{3.8}$$

Valuable sites are provided by the set of currently available proto-objects $\{\mathcal{O}_p(t)\}$ while the decision rule adopted depends on the current oculomotor state $z(t)$.

By assuming that the current oculomotor state is $z(t) = k$ and considering the conditional dependencies in the PGM of Fig. 3.2, Eq. 3.8 can be reduced to

$$\mathbf{r}_{new}(t+1) \sim P(\mathbf{r}_{FOA}(t+1)|\mathcal{O}(t), \boldsymbol{\xi}_k(t), \mathbf{r}_{FOA}(t)), \tag{3.9}$$

where $\boldsymbol{\xi}_k(t)$ are the most likely motor parameters for state $z(t) = k$, from which the angle and the amplitude of the gaze shift can be derived. Parameters $\boldsymbol{\xi}_k(t)$ and candidates $\mathbf{r}_{new}$ are obtained, at the simulation stage, via a stochastic sampling procedure. Indeed, stochastic sampling provides the computational tool to mimic human gaze shift variability (for details, see following Sec. 3.2 and [10] for an in-depth discussion).

At the most general level, if $z(t) = 1$ (saccade) has been chosen, then the expected reward of the shift $\mathbf{r}_{FOA}(t+1) \rightarrow \mathbf{r}_{new}(t+1)$ is computed with respect to the value of available proto-objects,

$$E\left[R_{\mathbf{r}_{new}}\right] = \sum_{p \in \mathcal{I}_V^k} \mathcal{V}(\mathcal{O}_p(t)) P(\mathcal{O}_p(t)|\mathbf{r}_{new}(t+1), \mathbf{T}). \tag{3.10}$$

In Eq. 3.10, the proto-objects $\mathcal{O}_p$ to be considered are those included in the set $\mathcal{I}_V^k$ of most valuable patches sampled from the whole image at time $t$, whose dimension is $|\mathcal{I}_V^k(t)| = N_V \leq N_p$. In Eq 3.10, $\mathcal{V}$ is the average value of proto-object $\mathcal{O}_p(t)$ with respect to the probability maps $P(\mathbf{V}_\ell(t)|\mathbf{R}(t))$.

Note that the set of proto-objects taken into consideration depends on index $k = z(t)$. If $z(t) = 0$, that is the eye is engaged in local exploration, then $\mathcal{I}_V^0$ restricts to the proto-objects localised within the current FOA area: thus, candidate point sampling occurs locally (fixational and small amplitude saccades).

Eventually, in either state, the next gaze location is determined so as to maximise the expected reward:

$$\mathbf{r}_{FOA}(t+1) = \arg\max_{\mathbf{r}_{new}} E\left[R_{\mathbf{r}_{new}}\right]. \tag{3.11}$$

The term $\arg\max_{\mathbf{r}_{new}}$ is the mathematical shorthand for "find the value of the argument that maximizes ...". In this instance, the argument is the next gaze candidate $\mathbf{r}_{new}$ and the expression to be maximised is the expected payoff.

It is worth recalling, from the discussion above, that what actually changes as a function of state is that, if the eye is feeding locally, and the task is a search task, then the effective reward $R\{P(\mathbf{r}_F OA(t), \mathbf{O})\}$ is computed through a "high resolution" detector/classifier. If the task is free-viewing then reward is obtained via $R\left(P(\mathbf{r}_F OA(t), \mathbf{L}(t)\right)$ computed on the high resolution priority map.

## 3.2 Simulation: gaze shift sampling

Here we provide details of a computational procedure to simulate the main features of the model and also we present some results by elucidating the whole computational process step by step; the corresponding representations that are obtained at the different levels of processing in the simulation are shown in Fig. 3.3. Following [10], we take the view that the gaze shift $\mathbf{r}_{FOA}(t) \rightarrow \mathbf{r}_{FOA}(t+1)$ is a way of sampling the visual landscape $\mathcal{W}(t)$ according to the current oculomotor action setting $\mathcal{A}(t)$ framed by the task $\mathbf{T}$.

**Pre-attentive representation** We assume that at $t = 0$, when then observer opens his eyes, a quick pre-attentive representation of the scene is made available [90]. To this end the fixation point $\mathbf{r}_{FOA}(0)$ is set at the centre of the picture, and the retinal image is simulated by blurring $\mathbf{I}$ through an isotropic Gaussian function centered at $\mathbf{r}_{FOA}(t)$, whose variance is taken as the radius of a FOA, $\sigma = |FOA|$, approximately given by $1/8 \min[w, h]$, where $w \times h = |\Omega|$, $|\Omega|$ being the dimension of the image support $\Omega$. This way we obtain the high resolution (HR) foveated image $\widehat{\mathbf{I}}_{HR}(0)$ (Fig. 3.3, top row, right picture); the foveated HR is mainly exploited to support local fixational movements and small saccades. This is then reduced through a pyramidal decomposition to $\widehat{\mathbf{I}}_{LR}(t)$, a low-resolution (LR) image mainly used during long relocation of the gaze. The foveation process will be updated for every gaze shift involving a large relocation, but not during fixational eye movements.

The LR image is adopted to roughly compute the initial feature likelihood $P(\mathbf{X}|\mathbf{FOA}, \mathbf{L})$. To such end, for what concerns face objects, we use the Viola-Jones detector by converting the AdaBoost outcome in a probabilistic output [8]. For what concerns textual objects, following [19] we simulate the localiser/detector using the text ground-truth. However, to be more realistic and compliant with the theoretical model, differently from [19], object likelihood is computed by using the output of Torralba's saliency [114] localised in the bounding box as given by the text region ground-truth. The motivation for this choice is that Torralba's saliency well correlates with text appearance [98] and it can be used as a rough but reliable estimate of its likelihood $P(\mathbf{FOA}|\mathbf{O} = text)$. Further, the main reason for using a simulated text likelihood estimator (instead of a real one such as in [27]) is that one can exploit *ad-hoc* control of the number of true positive / false positive regions. Having computed these coarse object-based maps it is easy to infer the initial priority map $P(\mathbf{L}|\widehat{\mathbf{I}}_{LR})$ [25] (Fig. 3.3, second row, left picture).

The value probability maps $P(\mathbf{v}_\ell(\mathbf{r}, 0))$ can be initialised as discussed in Sec. 3.1.2.1. One example, referring to the picture used in Fig. 3.3 is provided in Fig. 3.4. More in detail, such initialisation has been obtained through the following steps. At time $t = 0$,

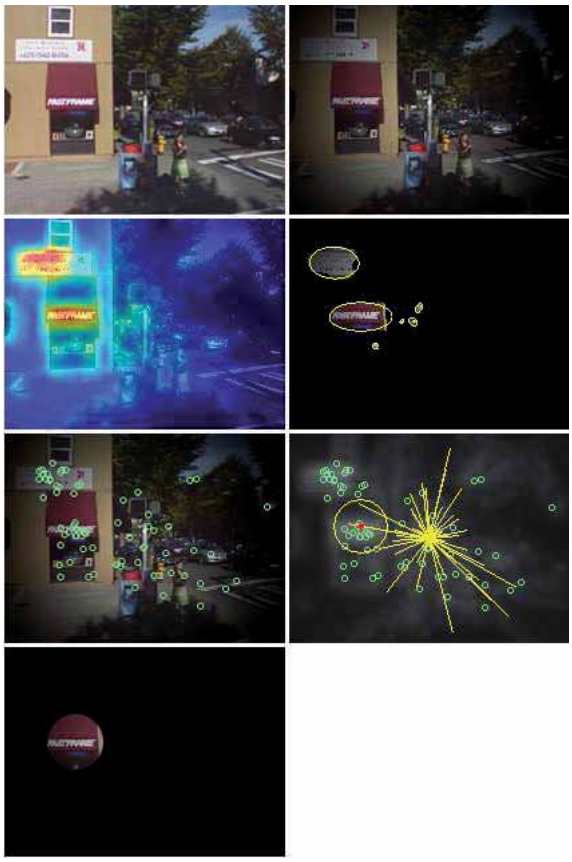


**Figure 3.3:** The main representations that are obtained at the different levels of processing in the simulation (details in the simulation discussion, Sec. 3.2). In this case the given task $\mathbf{T}$ is a "*Look for text regions*" task. From top to bottom, left to right: the original image $\mathbf{I}$; the foveated image $\widehat{\mathbf{I}}$ obtained by setting the initial FOA $\mathbf{r}_F OA(0)$ at the centre of the image; the priority map $\mathbf{L}$; selected proto-objects parametrised as ellipses $\theta_p(t)$; the interest points $O(t)$ sampled from proto-objects; the sampling process of candidate FOAs $\mathbf{r}_{new}(t+1)$ (Eq. 3.16) and the selection of $k$-th candidate point which maximises the expected reward $E\left[R_{\mathbf{r}_{new}}\right]$ (the big circles covers the points within $\mathcal{I}_V^k$); the sampled FOA $\mathbf{r}_F OA(t+1)$. All maps are depicted at the same resolution (HR) of the original image $\mathbf{I}$ for visualisation purposes. Value map initialisation follows the procedure illustrated in Fig. 3.4 below

(a) $\mathbf{T} = FV$: Text Value (b) $\mathbf{T} = FV$: Face Value

(c) $\mathbf{T} = S$: Text Value (d) $\mathbf{T} = S$: Face Value

**Figure 3.4:** The initial value probability maps $P(\mathbf{V}_\ell(0)|\mathbf{R}(0), \mathbf{T})$ calculated by weighting, at each spatial location, the estimated object maps (text and face) through the numerical payoff chosen for the given task $\mathbf{T}$ (see text for details). The input image is the one used for the example in Fig. 3.3. Free view (FV): $P(\mathbf{V}_{text}(0)|\mathbf{R}(0), FV)$ 3.4a and $P(\mathbf{V}_{face}(0)|\mathbf{R}(0), FV)$ 3.4b. Search for text (S): $P(\mathbf{V}_{text}(0)|\mathbf{R}(0), S)$ 3.4c and $P(\mathbf{V}_{face}(0)|\mathbf{R}(0), S)$ 3.4d. Probabilities, superimposed on the foveated image, have been scaled between $[0, 255]$ and colour coded, red colour denoting high probability, grey colour low probability.

the payoffs are set as a function of the task. We used $R_{text} = 50$ and $R_{face} = 100$ for $\mathbf{T} = FV$ (free-view), granting a higher attractiveness to faces with respect to text. For $\mathbf{T} = S$ (searching for text), $R_{text} = 100$ and $R_{face} = -50$. Then, the spatial feature map $P(\mathbf{X}(t)|\mathbf{L}(t), \mathbf{FOA}(t))$ computed for either $\mathbf{O} = face$ and $\mathbf{O} = text$ provides a pair of object likelihood maps that are used as approximate estimates of the object-based posterior density maps $P(\mathbf{L}(t), \mathbf{O} = face|\mathbf{T})$ (the posterior probability of observing a face object at a spatial location) and $P(\mathbf{L}(t), \mathbf{O} = text|\mathbf{T})$ (the posterior probability of observing a text object). Task $\mathbf{T}$ being assigned, the object maps are multiplied, with the payoff values chosen as above. To this point, the resulting maps are no longer probability maps. Thus, Eq. 3.2 is applied to each point of the maps for normalising between 0 and 1, and the task dependent value maps are eventually obtained, i.e. $P(\mathbf{V}_{text}(0)|\mathbf{R}(0), \mathbf{T})$, $P(\mathbf{V}_{face}(0)|\mathbf{R}(0), \mathbf{T})$. Such maps are shown in Fig. 3.4, where, for visualisation purposes, probabilities have been represented through colours. Note that, in order to fairly compare left and right probability maps, each colourbar at the right side of the map represents a colour (probability) range that is specific for that map. For instance the colourbar in Fig. 3.4c depicts the range $[130 = grey, \cdots, 255 = red]$, whilst the colourbar in Fig. 3.4c represents the range $[75 = grey, \cdots, 130 = red]$.

**Sparse representation of proto-objects:** Similarly to [10]) we make use of a sparse representation of proto-objects. Describing in terms of a foraging metaphor, the proto-objects can be conceived as foraging sites around which food items can be situated.

At any given time $t$, the foraging eye perceives a set $\mathcal{O}(t) = \{\mathcal{O}_p(t)\}_{p=1}^{N_P}$ of proto-objects or patches in terms of prey clusters, each patch being characterised by different shape and location. More formally, $\mathcal{O}_p(t) = (O_p(t), \Theta_p(t))$. Here $\Theta_p(t)$ is a parametric description of a patch, while $O_p(t) = \{\mathbf{r}_{i,p}\}_{i=1}^{N_{i,p}}$ is a sparse representation of patch $p$ as the cluster of interest points that can be sampled from it. More precisely, $\Theta_p(t) = (\mathcal{M}_p(t), \theta_p)$. The set $\mathcal{M}_p(t) = \{m_p(\mathbf{r}, t)\}_{\mathbf{r}\in L}$ stands for a map of binary RVs indicating at time $t$ the presence or absence of patch $p$. The overall map of proto-objects is given by $\mathcal{M}(t) = \bigcup_{p=1}^{N_p} \mathcal{M}_p(t)$. Here, $\mathcal{M}(t)$ is simply drawn from the priority map by deriving a preliminary binary map $\widetilde{\mathcal{M}}(t) = \widehat{m}(\mathbf{r}, t)\}_{\mathbf{r}\in L}$, such that $\widehat{m}(\mathbf{r}, t) = 1$ if $P(\mathbf{L}(t)|\widehat{\mathbf{I}}(t)) > T_M$, and $\widehat{m}(\mathbf{r}, t) = 0$ otherwise. The threshold $T_M$ is adaptively set so as to achieve 95% significance level in deciding whether the given priority values are in the extreme tails of the pdf. The procedure is based on the assumption that an informative proto-object is a relatively rare region and thus results in values which are in the tails of $P(\mathbf{L}(t)|\widehat{\mathbf{I}}(t))$. Then, following [122], $\mathcal{M}(t) = \{\mathcal{M}_p(t)\}_{p=1}^{N_P}$ is obtained as $\mathcal{M}_p(t) = \{m_p(\mathbf{r}, t)|\ell(B, \mathbf{r}, t) = p\}_{\mathbf{r}\in L}$, where the function $\ell$ labels $\widetilde{\mathcal{M}}(t)$ around $\mathbf{r}$ using the classic Rosenfeld and Pfaltz algorithm (implemented in the Matlab bwlabel function). We set the maximum number of patches to $N_P = 8$ to retain the most important patches. The patch map provides the necessary spatial support for a 2D ellipse maximum-likelihood approximation of each patch (see Fig. 3.3 second row, right picture), whose location and shape are parametrised as $\theta_p = (\mu_p, \Sigma_p)$ for $p = 1, \cdots, N_p$ (see [10] for a formal justification). Next, the procedure generates clusters of interest points, one cluster for each patch $p$:

$$O_p(t) \sim P(O_p(t)|\theta_p(t), \mathcal{M}_p(t) = 1, \mathbf{L}(t)). \tag{3.12}$$

By assuming a Gaussian distribution centered on the patch, Eq. (3.12) can be further specified as [10]:

$$\mathbf{r}_{i,p} \sim \mathcal{N}(\mathbf{r}_p; \mu_p(t), \Sigma_p(t)), \quad i = 1, \cdots, N_{i,p}. \tag{3.13}$$

We set $N_s = 50$ the maximum number of interest points and for each patch $p$, and we sample $\{\mathbf{r}_{i,p}\}_{i=1}^{N_{i,p}}$ from a Gaussian centered on the patch as in (3.13). The number of interest points per patch is estimated as $N_{i,p} = \lceil N_s \times \frac{A_p}{\sum_p A_p} \rceil$, $A_p = \pi\sigma_{x,p}\sigma_{y,p}$ being the area of patch $p$. Thus, the set of all interest points characterising the perceived scene can be obtained as $O(t) = \bigcup_{p=1}^{N_p} \{\mathbf{r}_{i,p}(t)\}_{i=1}^{N_{i,p}}$ (Fig. 3.3, third row, left picture).

**Determining the oculomotor action setting:** At the end of the proto-object sampling procedure we have at time $t$ the set $\mathcal{O}(t) = \{\mathcal{O}_p(t)\}_{p=1}^{N_P}$ of proto-objects in terms of interest points $O(t)$ , each patch being characterised by different shape and location, i.e., by proto-object parameters $\Theta_p(t)$. The first step is to determine the oculomotor state by sampling from the Bernoulli distribution via Eq. 3.7 with parameters determined by Eq. 3.6.

Assume that choice $z(t) = k$, with $k = 0, 1$, has been made. This allows to set the actual values of the motor parameters $\eta_k = \{\alpha_k, \beta_k, \gamma_k, \delta_k\}$. These are the parameters of the $\alpha$-stable distribution $f(\boldsymbol{\xi}_k; \eta_k(t))$, namely, the skewness $\beta$ (measure of asymmetry), the scale $\gamma$ (width of the distribution), the location $\delta$ and, most important, the characteristic exponent $\alpha$, or index of the distribution that specifies the asymptotic behavior of the distribution. The $\alpha$-stable distribution $f(\boldsymbol{\xi}_k; \eta_k(t))$ is then used to sample the stochastic components $\boldsymbol{\xi}_k(t) = \{\xi_{k,1}, \xi_{k,2}\}$ of candidate gaze shifts [10]:

$$\boldsymbol{\xi}_k(t) \sim f(\boldsymbol{\xi}_k; \eta_k(t)) \tag{3.14}$$

The $\alpha$-stable random vector $\boldsymbol{\xi}_k$ is sampled using the well known Chambers, Mallows, and Stuck procedure[22]. Here, parameters for longer shifts ($k = 0$) have been set to $\eta_0 = \{\alpha_0 = 1.6, \beta_0 = 1, \gamma_0 = 40, \delta_k = 200\}$ promoting a Lévy exploration, while for local walk ($k = 1$), $\eta_1 = \{\alpha_1 = 2, \beta_1 = 1, \gamma_1 = 22, \delta_1 = 60\}$.

**Deciding where to look next** Having determined the oculomotor action setting $\mathcal{A}(t)$, we can rewrite Eq. 3.9, that is the sampling of candidate gaze locations for the shift $\mathbf{r}_{FOA}(t) \rightarrow \mathbf{r}_{FOA}(t+1)$ as:

$$\mathbf{r}_{new}(t+1) \sim P(\mathbf{r}_{FOA}(t+1) | O(t), \theta(t), \boldsymbol{\xi}_k(t), \mathbf{r}_{FOA}(t)), \tag{3.15}$$

The shift is generated according to motor behaviour $z(t) = k$ and thus regulated by parameters $\eta_k$ conditioned on proto-objects sparsely represented through sampled interest points $O(t)$ and patch parameters $\theta(t)$.

Following the formal derivation of [10], exploiting the Euler-Maruyama discretisation of a Langevin-type stochastic differential equation, we sample $\mathbf{r}_{new}(t+1)$ by making explicit the stochastic dynamics behind the process as:

$$\mathbf{r}_{FOA}(t_{n+1}) \approx \mathbf{r}_{FOA}(t_n) - \sum_{p \in \mathcal{I}_V^k} \sum_{i \in \mathcal{I}_p} (\mathbf{r}_{FOA}(t_n) - \mathbf{r}_p(t_n))\tau + \gamma_k \mathbb{I} \tau^{1/\alpha_k} \boldsymbol{\xi}_k. \tag{3.16}$$

Thus the dynamics of gaze shift is determined by two terms. The first term $-\sum_{p \in \mathcal{I}_V^k} \sum_{i \in \mathcal{I}_p} (\mathbf{r}_{FOA}(t_n) - \mathbf{r}_p(t_n)$, is the deterministic drift that biases the walk towards the centre of gravity of selected interest points assuming that such *attractors* act as independent sources. Here $\mathcal{I}_p$ is the set of valuable interest points sampled from the patch $\mathcal{O}_p$ such that $p \in \mathcal{I}_V^k$ and $\tau = t_{n+1} - t_n$ is the integration time step.

The term $\gamma_k \mathbb{I} \tau^{1/\alpha_k} \boldsymbol{\xi}_k$ is the stochastic component which determines amplitude and orientation of the candidate gaze shift [10]. The symbol $\mathbb{I}$ denotes the $2 \times 2$ identity matrix and $\gamma_k$ the width of the $\alpha$-stable distribution from which $\boldsymbol{\xi}_k$ is sampled (Eq. 3.14). Notice that, due to the feed/fly switching of index $k = z(t)$ in Eq. 3.16, this random walk is a mixture of Lévy (large relocation) and nearly-Gaussian (local exploration) displacements.

Thus, Eq. 3.16 provides the explicit procedure for sampling candidate gaze shifts $\mathbf{r}_{FOA}(t) \rightarrow \mathbf{r}_{new}(t+1)$. Assume we sample $N_{new}$ such candidates, as shown in Fig. 3.3, third row, right picture. Then the decision to saccade is taken in order to maximise the expected reward of having valuable interest points in the neighbourhood of the

candidate shift (represented in the same picture as a wide yellow circle). This can be obtained by writing Eq. 3.10 as

$$E\left[R_{\mathbf{r}_{new}}\right] = \sum_{p \in \mathcal{I}_V^k} \sum_{i \in \mathcal{I}_p} \mathcal{V}(\mathbf{r}_{i,p}(t)) \mathcal{N}(\mathbf{r}_{i,p}(t) | \mathbf{r}_{new}(t+1), \Sigma_s). \tag{3.17}$$

Finally, the actual gaze shift is obtained through Eq. 3.11 (Fig. 3.3, bottom picture)

Recall from Secs. 3.1.2.3 and 3.1.2.1 that in the feeding state we have to compute the effective reward. In particular, if the task is a search task, we stated that the effective reward $R\left(P(\mathbf{r}_F OA(t), \mathbf{O})\right)$ should be computed through a "high resolution" detector/classifier. To such end, if the object to look for is a face we use the probabilistic version of the Viola-Jones detector, but working on the HR image (which entails higher precision); if we are searching for text, we straightforwardly use the HR text ground-truth as a "perfect classifier" (oracle). To complete the picture, at each shift the IOR is simulated on the priority map by applying an inverse Normal suppression function at $\mathbf{r}_F OA(t)$, as in [108].

All parameters of the model have been tuned by using a subset of 50 images from the Microsoft dataset and related eye tracking data (see Secs. 4.1, 4.3).

Finally, in order to get a better understanding of the inner workings of the model, we show an example where we successively switch off the different control levels. Results are shown in Fig. 3.5. The top row presents two scan paths obtained assigning the task of "*Look for text*" (left picture) and the task of "*Look for people*" (right picture); at this level the simulation of the model is working in full mode. The central row presents results obtained when no task is given and control by value and payoff is inhibited. The left picture shows the priority map after the first central fixation. In this case, $\mathcal{W}$ relies entirely on the priority representation and the on proto-objects that can be sampled from it; also, the prior probability of objects given the task, $P(\mathbf{O}|\mathbf{T})$, is taken as a uniform distribution and hence the contribution by early saliency becomes stronger. The forager's willingness to feed or to fly $\mu(R(\mathbf{r}_F OA(t)))$ (Eq. 3.6) is set to a constant, and the decision rule in Eq. 3.17 is simplified by letting $\mathcal{V}(\mathbf{r}_{i,p}(t)) = \mathbf{r}_{i,p}(t)$, that is $\mathcal{V}$ is to be considered an identity function, since value plays no role at this stage. The right picture on the same row depicts one simulated scan path where the central bias effect of the foveated priority map is readily apparent. The bottom row shows the simulation of the model when no object information is available, thus $P(\mathbf{FOA}(t)|\mathbf{O}) = P(\mathbf{FOA}(t))$ and the gaze shift process (right) only nourishes on early saliency yet modulated by foveation (left).

## 3.3 Conclusion

Let us summarize the main points tackled in this chapter. We proposed a human-like visual attention model aiming to describe the mechanism behind the attention allocation in street view images. We specifically address the point of how to model some top-down constrains and how they could interact with stimulus driven saliency. Some progress in this direction have been recently made by approaches aiming to embed object detection in the saliency description, and the approach of Torralba accounting for the Gist.

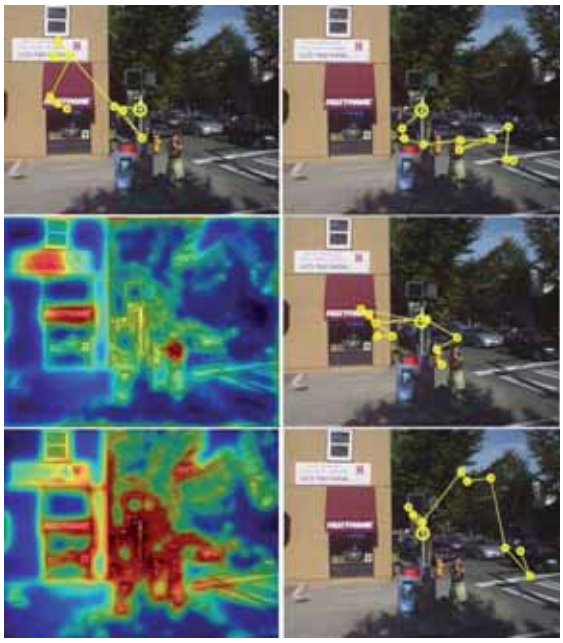


**Figure 3.5:** Inhibition of levels of representation and control. Top row: scan path generated when the given task **T** is "*Look for text*", similarly to Fig. 3.3 (left); scan path generated when the model simulates a "*Look for people*" task (right). Middle row, no task and value assigned, but object likelihood is still computed: the foveated priority map **L** (left, red colour coding for high priority locations, blue for low priority) and one generated scan path (right). Bottom row: when the object likelihood is not computed, the priority map collapses to a classic early saliency but modulated by foveation (left); a corresponding scan path (right). All maps are depicted at the same resolution (HR) of the original image **I** for visualization purposes

Largely inspired by the psychological model of Rensink and the discussion provided in Chapter 2, here we propose a computational model describing attention deployment as coming from the interaction between an attentional process and a motor process. On a side the attentional process assigns value to proto-object like sources of information and on the other side the motor process implements a selection mechanism based on human like oculomotor biases.

The model is described in terms of Probabilistic Graphical Model, a graph-based representation where nodes denote RVs and arrows code conditional dependencies between RVs. The structural dependencies shaping such hierarchy can be formalised making explicit the distributions at the different levels of representation: the object and object feature level, the priority map level, the priority map level, the local feature level that ties object features to prioritized locations, the foveated image level. Given the complexity of the model we provide details of a computational procedure to simulate the main features of the and also we present some results by elucidating the whole computational process step by step.

In our formulation the choice of where to look next is task-dependent and oriented to classes of objects embedded within the picture. The dependence on task is taken into account by exploiting the value and the reward of gazing at certain image patches or proto-objects that provide a sparse representation of the scene objects.

# Chapter 4

# Experimental Evaluation of the Model

In this chapter we confront the scan paths produced by the model with those from eye-tracked human subjects. Such comparison is qualitative in terms of observable scan paths and quantitative in terms of statistical similarity of oculomotor behaviour. Gaze shift amplitude distributions of human observers are compared to those obtained from simulations by means of studying the amplitude distribution [110, 112], and in particular of the corresponding complementary cumulative distribution function, following the standard convention in the literature [10].

Comparative experiments are performed using eye tracking data from both a publicly available dataset of face and text and from newly performed eye-tracking experiment on a complex dataset of street view pictures containing text. In both cases data are collected by a video based desktop eye tracker.

In order to provide quantitative results concerning semantic aspects of the text search, we also analyse the discriminability performance of simulated scan paths, in the specific case of the "*Look for text*" task, in terms of average True Positive Rate and False Positive Rate. For all quantitative assessments we used as a baseline control model, the Itti & Koch model as implemented in the latest version of the saliency tool box downloaded from the saliencytoolbox web page.[1]

## 4.1 Datasets

**Cerf's Fixations In FAces dataset.** This dataset[2] contains Faces a subset of 229 images ($1024 \times 768$ pixels) showing frontal faces in various sizes, locations, skin colours, races, etc. Each image has a corresponding background image with no faces for comparison. The dataset includes the fixations data recorded via eye-tracking experiments of 8 subjects in free viewing conditions (see [19] for details). Dataset also provides annotation of faces position in the form of bounding box coordinates.

[1] http://www.saliencytoolbox.net
[2] http://www.fifadb.com/

**Epshtein's Microsoft dataset.** This publicly available dataset[3], consists of 307 colour street view pictures of sizes ranging from $1360 \times 1024$ to $1024 \times 768$ pixels. The text content is embedded in the scene in the form of shop names, street signs or advertisements and it is usually not located at the centre of the image, nor covering a large region of the image, so as to make the localisation problem more difficult (see [3] for details). We use this dataset for specifically assessing the difference between human subjects and model's simulation behaviour when looking at street view pictures containing text objects.

## 4.2 Experiment 1

The aim of this experiment was to compare the motor behaviour predicted by the model with experimental scan paths from human subjects in free viewing condition ($\mathbf{T} = FV$). For this experiment we used the Fixations In FAces dataset. Pictures contained either faces, or text regions or both.

First comparison is qualitative: we generated 20 scan paths for each image and compared them to those exhibited by human observers by choosing the most similar scan paths in terms of fixations coordinates, duration, and time occurrence. Some typical results obtained are presented in Fig. 4.1, in which a face is present on the left part of the scene 4.1a and when the face is removed 4.1b. For both the human observer (scanpath on the left side, in red) and the model (scanpath on the right side, in yellow) fixations goes to the face, as long a face is present in the scene. Similarly in Fig. 4.2 showing a scene in which face and text are both present 4.2a and when the face is removed and the only text appear 4.2b, we observe that face and text are both attracting fixations even if the task is a free viewing task. In the second case, when the face object is missing, fixations goes to other semantically important object of the scene. The model, that is including partial knowledge of the objects present in the scene as described in Chapter 3, is than able to mimic observer's oculomotor behaviour in free viewing observation of complex street view images.

More quantitatively, we studied the empirical distributions of gaze shift amplitudes [112, 110, 10] by analyzing eye-tracking data. To this end the gaze shift samples from all the experiments, regardless of the observers, are aggregated together and used in the same distribution. Aggregating the data is a reasonable assumption because every scanpath obtained from the same image is subject to the same or similar visual constraints and the same technique is used in other studies of Levy walks (e.g., [94]) but also in eye-tracking experiments [110]. For a precise description of the tail behaviour, the laws governing the probability of large shifts, we account for the upper tail, or complementary CDF (CCDF) of jump lengths, following the standard convention in the literature. The upper tail of the distribution of the gaze shift magnitude $X$ can be defined as $\overline{F}(x) = P(X > x) = 1 - F(x)$, where $F$ is the cumulative distribution function (CDF).

We introduce a control condition, running simulations were virtual observers are 'viewing' the same set of images. We used as baseline control model the Itti *et. al* model [53]. For each image, the virtual observer made the same number of simulated

[3] http://research.microsoft.com/en-us/um/people/eyalofek

(a)

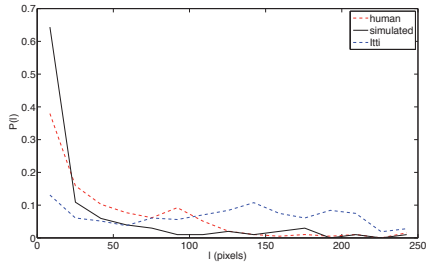
(b)

**Figure 4.1:** Scan paths generated while free viewing a picture from Fixations In FAces dataset, when a face is present 4.1a and when the face is removed 4.1b. Left (in red colour), scan path obtained eye-tracking a human observer; right, model output (in yellow)

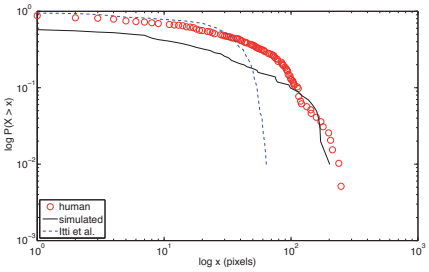
(a)

(b)

**Figure 4.2:** Scan paths generated while free viewing a picture from Fixations In FAces dataset. In 4.2a face and text are both present, whilst in 4.2b the face is removed. Left (in red colour), scan path obtained eye-tracking a human observer; right, model output (in yellow)




(a) Gaze shift amplitude distribution




(b) CCDF

**Figure 4.3:** Comparing the oculomotor behaviour generated by humans with either the one simulated by the proposed model and by the one of Itti model. The comparison is provided in terms of gaze shift amplitudes on the Fixations In FAces dataset . Top panel (4.3a) shows the empirical distributions of gaze shift amplitudes; bottom panel (4.3b) shows the double log-plots of the corresponding CCDFs.

saccades as the human observer had on that scene. Comparison of the oculomotor behaviour generated by humans with the one simulated by the proposed model and by the model of Itti are illustrated in Fig. 4.3. Top panel (4.3a) shows the empirical distributions of gaze shift amplitudes; bottom panel (4.3b) shows the double log-plots of the corresponding CCDFs. It can be noticed that Itti *et al.* model does not show the characteristic positively skewed distribution of gaze shift amplitudes exhibited by human scan paths and well captured by the proposed model. Differences in gaze shift statistics can be easily appreciated from the CCDF plot (Fig. 4.3b), in this regard the tail behaviour of the gaze amplitude distribution of the proposed model fits well the human. The tail behaviour of the Itti model stays far from both human and proposed model distribution. These results are consistent with results presented by Tatler *et al.* [110].

The fit between the empirical distributions of eye-tracked and simulated gaze shifts amplitudes, is also assessed via the two-sample Kolmogorov-Smirnov (K-S) test, which is very sensitive in detecting even a minuscule difference between two populations of data, and the standard Mann-Whitney U (MWU) test, to assess the null hypothesis that two samples have the same median (central tendency). All tests are performed at the level of significance $\alpha = 0.05$ and repeated for ten model simulation trials.

According to the K-S test, the simulated distribution was shown to be not significantly different from the human one for 70% of cases (average value for all trials). MWU assessed the same central tendency for 92% of cases. The control model always fails both tests.

## 4.3 Experiment 2

With this experiment we aim to assess the difference between human subjects and model's simulation behaviour when looking at street view pictures containing text objects. We use the Microsoft dataset, which comprises images more complex than those in the Fixations In FAces dataset, as it is mostly adopted for "text-in-the-wild" detection/classification contests. This dataset offers the advantage of having publicly available ground-truth for text regions but no eye-tracking data is available. Thus we conducted some eye tracking experiments to collect data in both free viewing and text search condition.

### 4.3.1 Eye Tracking data collection

The eye-tracking experiments have been conducted using a video-based SMI RED eye tracker (SensoMotoric Instruments, Teltow, Germany) at a sampling rate of 120Hz., with automatic head movement compensation (tracking range, $40 \times 30$ cm at 70 cm distance). The infrared video-based system has an instrument spatial resolution of $0.03°$ and an absolute gaze position accuracy of up to $0.4°$.

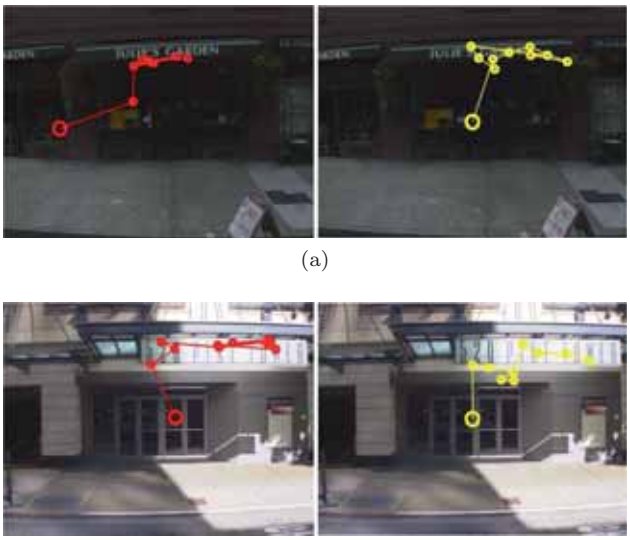
The experiment took place in a dimly lit room in the Computer Vision Center in Barcelona. Subjects were seated in a contact-free setup, 70 cm in front of a 22-inch LCD monitor (60 Hz refresh rate, 58.18 dpi). Stimulus resolution was $1024 \times 768$ pixels at both sites and subtended approximately a visual angle of $36.6°(w) \times 27.4°(h)$. A 9-point calibration of the eye tracker was carried out at the onset of every trial.

**Participants recruitment:** Two groups of six naive adults (3 women and 3 men, composing the first group, 2 women and 4 men for the second group, range 25-44 years, mean 32 years) participated in the experiment. All participants were native speakers of Spanish and had normal or corrected to normal vision.

**Task instructions:** Each subject was asked to look at pictures presented on the monitor. Two tasks were considered. A search task, $\mathbf{T} = S$, formulated in terms of "*Look for text regions within the pictured scene*" was assigned to the first group; a free-view task, $\mathbf{T} = FV$, formulated as a generic "*Guess the city from the pictured scene*," so as to motivate the participants, was given to the second group. Pictures were presented in randomized order and each picture was shown for 5 seconds. Stimulus luminance was linear in pixel values.

### 4.3.2 Comparison

Qualitative comparison are performed as in Experiment 1. Some examples representative of scanpaths obtained for the Task $\mathbf{T} = S$ are provided in Fig. 4.4 when text is the main semantic object, and in Fig. 4.5 when other semantic objects (face, people)



(a)

(b)

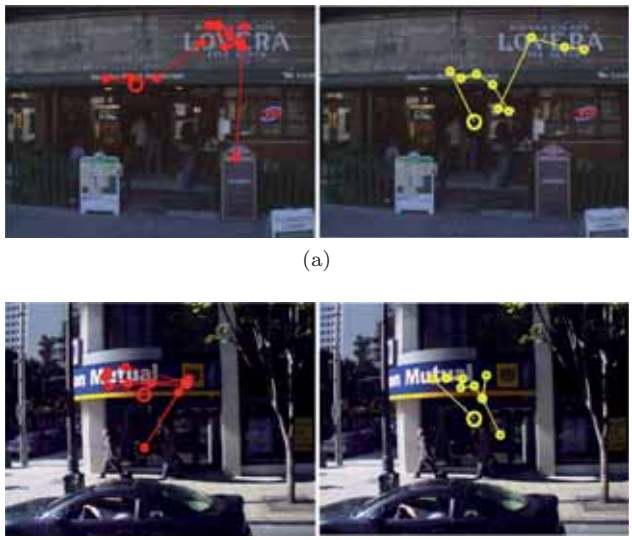
**Figure 4.4:** Scan paths generated under the "*Look for text regions*" task for pictures from the Microsoft dataset, where text is the main semantic object class. Left (in red colour) the scan path obtained from eye-tracking a human observer; Right (in yellow colour) the simulated scanpath from our model.

appear in the picture together with the text. In both the cases the scan path obtained from our model (on the right in yellow colour) is able to mimic well observer's scanpath as recorded from eye-tracking experiments (on the left side in red colour).

We also observe that the scanpath goes on the bigger text objects as it was expected, as long as the text is always attractive for fixations [19, 124]. It is worth noting that in the Microsoft dataset the vast majority of images contains text regions as the most semantically relevant objects appearing within the image scene. This result is also reflected in the cumulative statistics of shift amplitudes, which result to be fairly similar for both text search and free viewing Task, as it can be appreciated at a glance from Figs. 4.7 and 4.8 below.

Fig. 4.7a and Fig. 4.8a shows the empirical distribution of gaze shifts amplitude for respectively $\mathbf{T} = S$ and $\mathbf{T} = FV$ . The distribution for the Itti model is the same in both the plots as long that model does not allow to condition the scanpath in the task. The human distributions of gaze shifts are very similar although coming from separate eye tracking experiments under different experimental conditions as described before. The distribution coming from our model is also very similar in both the task and similar to the human distribution.

Nevertheless, there are cases that bear a specific interest. For instance, we show one such example in Fig. 4.6. This can be considered as the "dual" of the example provided in Fig. 4.2 in which, under the unchanged task ($\mathbf{T} = FV$), one class of objects was removed and specifically face and text were both present in one picture and only text was present in the other. Here instead both classes of objects ($\mathbf{O} = face$



(a)

(b)

**Figure 4.5:** Scan paths generated under the *look for text* for pictures from Microsoft dataset when other semantic objects (faces, people) are embedded in the picture together with text. Left (in red colour), the scan path obtained from eye-tracking a human observer; Right (in yellow colour) the simulated scanpath from our model.

and $\mathbf{O} = text$) are retained, but the task is switched from $\mathbf{T} = S$ (4.6a) to $\mathbf{T} = FV$ (4.6b). It can be noted that for $\mathbf{T} = S$ (4.6a, left), the girl is treated as a "distractor" by the human observer, whilst for $\mathbf{T} = FV$ (4.6b) it is competing for attracting gaze though being less visible and physically salient with respect to text regions in the scene (4.6b). The model achieves a similar behaviour by the different assignment of value in either task (cfr. Fig. 3.4).

Under "*Look for text regions*" task, by performing the K-S test as in the previous experiment, the simulated distribution resulted no significantly different from the human one for an average 79% of cases. MWU assessed the same central tendency for 89% of cases. For the task "*Guess the city*", the K-S test found no significant differences between the two distributions 89% of cases. MWU assessed the same central tendency 96% of cases.

## 4.4 Discriminability performance in Text Search

In order to provide quantitative results concerning semantic aspects that the "*Look for text* " task brings in, we have performed the following analysis. Since the Microsoft dataset includes the maps of text objects located in each image we can compute the ground-truth binary text map $\mathcal{TM}$ with $\mathcal{TM}(x, y) = 1$ for pixels $(x, y)$ belonging to target objects, $\mathcal{TM}(x, y) = 0$ for points outside text regions. Given the $s$-th scan path on the same image, we obtain the binary fixation map $\mathcal{FM}_s$ by considering the first 10 fixations of $s$ and by setting to 1 points within the circular region defining around

(a) "*Look for text* "

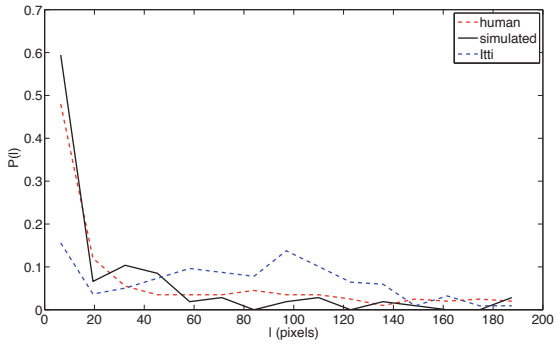
(b) "*Guess the city*"

**Figure 4.6:** Scan paths generated under the "*Look for text* " task, for a picture where other semantic objects (faces, people) are embedded in the picture 4.6a and under the "*Guess the city*" task, 4.6b. Left (in red colourrealis), scan path obtained eye-tracking a human observer; right, model output (yellow)

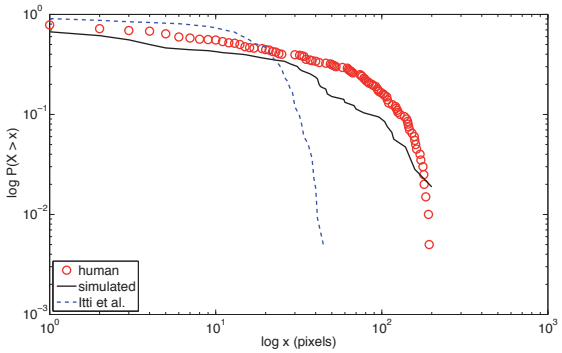
each fixation point, and to 0 points outside such areas. For what concerns the radius of each fixational region, we set $\varphi = 2°$ of visual angle. The size of this "functional fovea" is slightly larger than the 2° window spanned by a fixational eye movement [47, 107] and corresponds to the $7° - 8°$ window that can be searched effectively in one fixation [41]. Yet, it is smaller than the conservative estimate by Shioiri and Ikeda, who define 10° of visual angle the maximal window over which high-resolution pictorial information can be extracted [99]. By taking into account the experimental viewing conditions (viewing distance $v_d = 70$ *cm*, screen resolution $s_r = 58.18$ *dpi*), the radius $\varphi$ of region can be calculated in pixel units as

$$r_{fix} = \varphi \frac{1}{2\tan^{-1}\left(\frac{1}{2v_d}\right)} \frac{\pi}{180} \frac{s_r}{2.54} \quad (pxl). \tag{4.1}$$

Thus, $r_{fix} \approx 55$ pixels. The reason for considering a small circular region circumscribing a fixation rather than simply the fixation point itself is either to account for the fixational movement and to provide a different weight for fixations falling in the neighbourhood of object border with respect to fixations occurring within object. Then, for each scan path $s$, we can measure the True Positive Rate, $TPR_s = |TP_s|/|P|$ and the False Positive Rate, $FPR_s = |FP_s|/|N|$, where $|P|$ is the number of points within the object set,
$P = \{\mathcal{TM}(x, y)|\mathcal{TM}(x, y) > 0\}$ and $|N|$ is the number of points outside. The true positives and false positives, $|TP_s|$ and $|FP_s|$, respectively, are determined by counting the non zero points of the sets

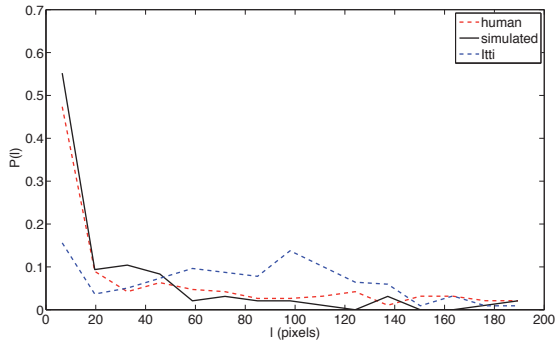



(a) Gaze shift amplitude distribution

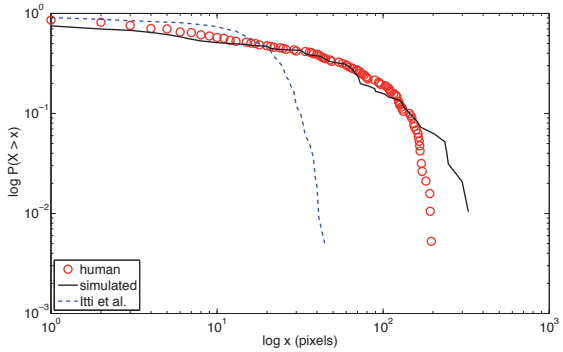



(b) CCDF

**Figure 4.7:** Comparing the oculomotor behaviour generated by humans and simulated by the model on the Microsoft dataset in terms of gaze shift amplitudes. The task was "*Look for text regions*". Top panel (4.7a) compares the empirical distribution of gaze shift amplitudes; bottom panel (4.7b) shows the double log-plot of the corresponding CCDF.




(a) Gaze shift amplitude distribution




(b) CCDF

**Figure 4.8:** Comparing the oculomotor behaviour generated by humans and simulated by the model on the Microsoft dataset in terms of gaze shift amplitudes. The task was "*Guess the city*". Top panel (4.8a) compares the empirical distribution of gaze shift amplitudes; bottom panel (4.8b) shows the double log-plot of the corresponding CCDF.

| Observers | $TPR$ | $FPR$ | $d'$ |
|---|---|---|---|
| Humans | $0.511 \pm 0.075$ | $0.057 \pm 0.008$ | 1.60 |
| Model | $0.351 \pm 0.091$ | $0.052 \pm 0.009$ | 1.21 |
| Control model | $0.129 \pm 0.17$ | $0.079 \pm 0.007$ | 0.27 |

**Table 4.1**
$TPR$, $FPR$ AND $d'$ FOR OBSERVERS AND VIRTUAL OBSERVERS SIMULATED BY THE PROPOSED MODEL AND BY THE CONTROL MODEL

$$TP_s = \mathcal{TM} \bigcap \mathcal{FM}_s, \quad FP_s = \mathcal{TM}^{\complement} \bigcap \mathcal{FM}_s, \tag{4.2}$$

where $\mathcal{TM}^{\complement}$ is the complement of the binary map $\mathcal{TM}$. Then, the average $TPR_s$ and $FPR_s$ are calculated taking into account all the scan paths generated within each group of observers: human, model and control model. The final total averages $TPR$ and $FPR$ computed on all the images of the dataset for each group are reported in Table 4.1, where the performance of the proposed model can be compared with human and control model performance. As previously, the Itti *et al* model was used as a baseline control model.

It can be seen from Table 4.1 that the average sensitivity ($TPR$) - in our case the average proportion of actual positives (pixels belonging to text regions) that have been correctly spotted within the first 10 fixations - is similar in both human and model generated scan paths, while the control model exhibits a lower sensitivity. Analogously, humans and model are close in terms of specificity ($1 - FPR$), at variance with the control model, which is characterized by marginally lower specificity. These results are statistically significant as it can be seen by measuring the difference between the spotting error rate of human observers and the error rate of a model $m$ (either the proposed or the control model). This way, the statistic $z_{obs,m} = (p_{obs} - p_m)/\sqrt{2p(1-p)/n}$ [102] is obtained, with $p = (p_{obs} + p_m)/2$, $n = |N| + |T|$, and where $p_{obs}$ and $p_m$ are the proportions of test samples (pixels) incorrectly spotted by observers and the model $m$ respectively. The statistic has a standard normal distribution [102], and the null hypothesis that human subjects and the model have the same error rate cannot be rejected ($|z_{obs,model}| = 0.07 < Z_{0.975} = 1.96$, two-sided test, $p = 0.94$, significance level $\alpha = 0.05$); conversely, the difference between the control model and humans is remarkable ($|z_{obs,control}| = 70.5 > Z_{0.975}$, $p < 0.001$). The same conclusion is achieved via McNemar's chi-square test [34], with Yates' correction ($p = 0.97$ and $p < 0.001$, respectively, $\alpha = 0.05$).

Similar results are obtained by computing, as index of performance, the discriminability $d'$ (cfr. Table 4.1), which summarises the capability of the scan path to separate text objects and non text regions, regardless of the statistical decision criterion. This index was calculated as $Z_{TPR} - Z_{FPR}$, where $Z_{TPR}$ is the $z$-transformed $TPR$ and $Z_{FPR}$ is the $z$-transformed $FPR$.

## 4.5 Conclusion

In this chapter we confront the scan paths produced by the proposed model with those from eye-tracked human subjects and a control model. Performance were assessed both in qualitative terms observing some sample scanpath and quantitative in terms of statistical similarity of oculomotor behaviour.

We used first a publicly available dataset providing pictures and eye tracking data to test our model capability to simulate the human behaviour. It can be noticed that Itti *et al.* model does not show the characteristic positively skewed distribution of gaze shift amplitudes exhibited by human scan paths and well captured by our proposed model. Differences in gaze shift statistics can be easily appreciated from the CCDF plot (Fig. 4.3b), in this regard the tail behaviour of the gaze amplitude distribution of the proposed model fits well the human. The tail behaviour of the Itti model stays far from both human and proposed model distribution. These results are consistent with results presented by Tatler *et al.* [110]. According to the K-S test, the simulated distribution resulted no significantly different from the human one for 70% of cases (average value for all trials). MWU assessed the same central tendency for 92% of cases. The control model always fails both tests.

We use a dataset of street view picture for specifically assessing the difference between human subjects and model's simulation behaviour when looking at more complex pictures containing text objects. We recorded eye tracking data on this specific dataset to explicitly look at two different working conditions: under a free viewing task and under a text search task.

In conclusion our model is able to describe the statistical properties of human gaze shifts, mimicking well the observer's oculomotor behaviour in the observation of complex street view images. In the limitation of our analysis to the gaze shift amplitudes, our model is approximating well the human behaviour in both free viewing and text search task. Regarding the only text search task we assessed the model performance in terms of discriminability $d'$ metric, summarising the capability of the scan path to separate text objects and non text regions. Our model scores closer to the human performance than the one of Itti in terms of $d'$ metric.

# Chapter 5

# Experimental evaluation in outdoor settings

In this chapter we study to what extent the baseline attentive model presented in Chapter 3 can account for human eye movements in an experimental setting by far more complex than the laboratory setting used in the experimental evaluation discussed in Chapter 4. We make use of a mobile eye tracker to record human eye movements in free condition, allowing head movements and small body rotational movements, while subjects are looking for text object in 360 degrees directions, in an outdoor recording setting. By means of this experimental evaluation we are interested to see whether the proposed model can account for human eye movements in a close to real world setting. Variations to the baseline model are made on the basis of oculomotor biases learned from eye tracking data of human subjects. The scanpaths produced by the different sets of models are compared one with the other by employing a definition of performance accounting for how accurate are the fixations in hitting the in-scene text targets. Human performance and the chance level are measured using the same definition of performance, and are used to define the upper and lower bound of the model performance.

## 5.1 Challenges in outdoor locations

The use of a mobile device in outdoor locations brings new challenges, some due to the use of a mobile device others due to the semi controlled environment we run experiments in. Planning the mobile eye tracking experiments we accounted for: scene, subject and recording device. Differently from desktop eye tracker setups, where the stimulus is typically a picture or a video, when doing mobile eye tracking the stimulus is a piece of the real world and we will be taking multiple views of it over the time, one for each eye tracking recording session. Given such settings we have to account for some amount of variability in the scene, thus some variability into the stimulus provided to each subject under test, and experiments have to be designed accordingly, to minimize such variability.

The scenes of interest in this research are limited to street view scenes, typically

commercial areas in which we can expect to find some text in the form of shop names and signs. Perfect candidates to this objective are city center squares as well as commercial streets. Accounting for scene changes in the time range from few hours to few weeks, we can roughly distinguish between target objects that are supposed to be stable, such as buildings, shops names, city signs and objects, such as people and cars, that are supposed to move. In this research we assume that all the text objects (targets / objects of interest) are not moving (and stable) across all eye tracking experiments. Note that this assumption is not always true because in real life scenarios new text panels might appear from day to day; to this end the best we could do was to choose locations where the text content was not expected to change position or appearance in the time frame of a few weeks. We also assume that all the other objects allowed to move will not contain text, and play the role of distractors.

Illumination conditions are strongly affecting the visibility of the scene. Running an outdoor experiment at different times of the day may strongly affect the scene appearance, making text more or less readable. In pilot experimentation we observed the extreme cases of sun light casting shadows on the text, as well as masking text by making reflections on glass or reflective surfaces. We thus planned to perform experiments so that we minimise illumination changes in the scene.

Mobile eye tracking experiments make the data collection a much more slower and time consuming activity than traditional desktop eye tracking experiments due to several factors: 1. the overall experiment time is longer due to the need of moving to the different locations, 2. subjects cannot be queued nor allowed to participate more than one at a time, to avoid biasing 3. experiment scheduling has to account for weather conditions, 4. number of experiments per day is also limited by the need of running experiments in specific time of the day, (illumination variation during the day) Running experiments we also observed that is best to run experiments with subjects not using corrective glasses, due to technical difficulty in pupil tracking.

Mobile eye tracking devices differ from the desktop ones for recording information in a per frame reference system. As lacking of a fixed and predefined two-dimensional reference system, the mobile eye tracking is making the eye movements analysis more difficult than the desktop setting, and usually calling for manual annotation of video sequences.

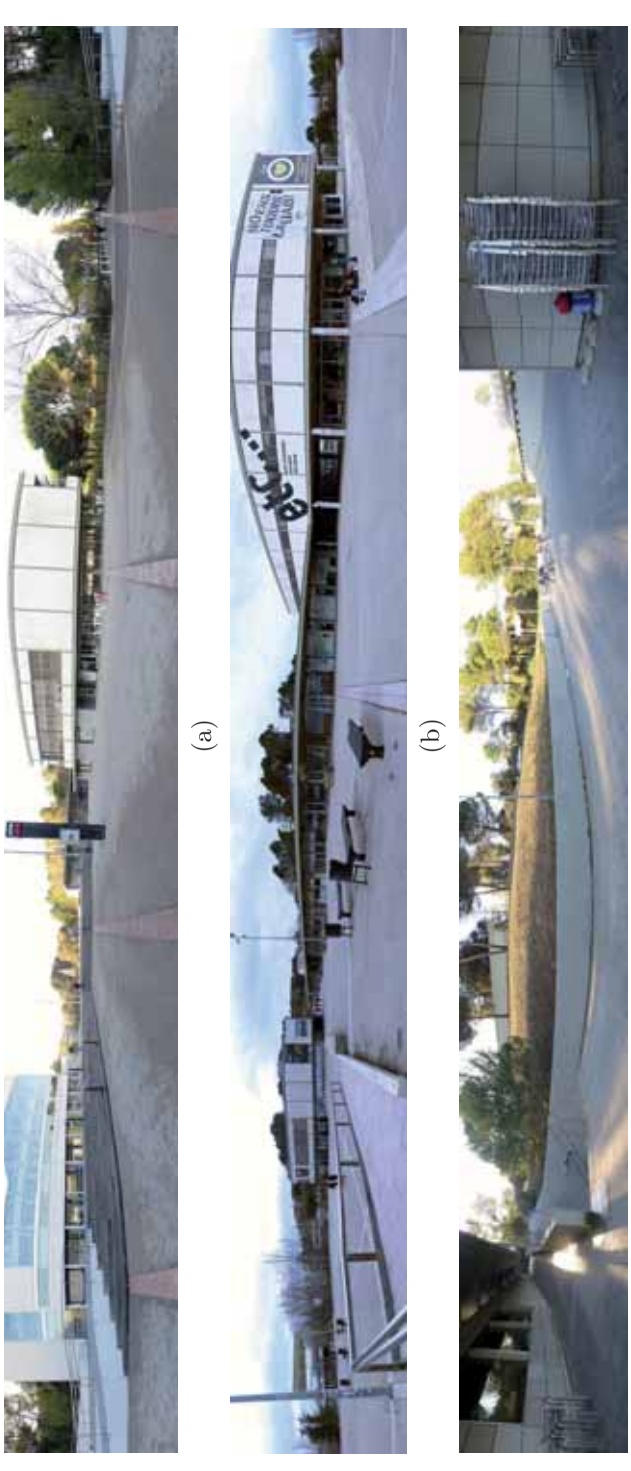


**Figure 5.1:** Panoramic pictures of the three locations from mobile eye tracking experiments: (a) Location n.1, (b) Location n.2, (c) Location n.3

### 5.1.1 Eye Tracking Glasses

Technological development made available a large number of devices for mobile eye tracking from several manufacturers as SMI Glasses, Tobii Glasses, Mangold Mobile Eye and ASL Mobile Eye-XG. For this research we made use of the SMI Eye Tracking Glasses 1.0 device, equipped with a front camera to record the scene at 1280x960 pixel resolution and 24 fps from the subject's head point of view, as in Fig. 5.2.

A binocular system to record eye movements consisting of two cameras and infra red illuminators. Movements of both eyes are captured by two small cameras on the rim of the glasses, while automatic parallax compensation ensures accurate data over all distances with no need for manual adjustments. Table 5.1 reports technical data.


**Figure 5.2:** The mobile eye tracking glasses used to record eye movement. Front, left, top views. Images taken from the ETG SMI user manual.

| | |
|---|---|
| Eye tracking principle: | Binocular eye tracking with automatic parallax compensation; Pupil/CR, dark pupil tracking |
| Temporal resolution: | 60Hz and 30Hz binocular |
| Gaze position accuracy: | 0.5 ° over all distances, parallax compensation |
| Gaze tracking range: | 80 ° horizontal, 60 ° vertical |
| HD scene camera, resolution: | 1280x960p at 24 fps; 960x720p at 30 fps |
| HD scene camera, video format: | H.264; Field of view: 60 ° horizontal, 46 ° vertical |

**Table 5.1**
SMI Eye Tracking Glasses, device technical data.

## 5.2 Experimental design

Mobile eye tracking methods provide a good solution for studying perception in variable contexts and a wider degree in the freedom of movement, than traditional desktop eye tracking experiments. However such freedom brings new challenges as observed in Sect. 5.1. We devised the following experimental settings to allow both a good freedom of movement to the subjects under test and, at the same time, keep the data analysis processing as simple as possible:

**Method:** Subjects were instructed to fully observe the scene in 360 degrees from a single observation point. Observers' movements were restricted by asking them to do not walk and do not move from the assigned point. Subjects were allowed to only turn around by taking small steps on the place. Imposing the subject to do not move across the scene allows us to describe the scene as an unfolded panoramic picture. Sect. 5.3 describes the procedure we use to obtain the mapping between eye tracker camera scene view and the panoramic scene.

**Locations selected:** We selected three locations inside the university campus. We opted for safe pedestrian areas at the commercial area of the campus close to shops, restaurants, library and train station. Panoramic views of the locations are shown in figure 5.1

**Participants recruitment:** We recruited 14 participants and for all of them performed the eye tracking experiments at all the selected locations. The participants were students and researchers recruited at the science department, half native Spanish and half non-Spanish and English speakers. All able to read the text without use of corrective glasses. Participants were compensated for their time.

**Task instructions:** Subjects were instructed about the task by means of written instructions. The task instructions were provided before starting the real experiment to ensure the subject understood the task and understood the importance of not moving from the assigned place. The written instructions were as follows: "You are allowed to look around in 360 degrees without moving from the assigned point. Take all the time you need to observe the scene but look at it fully and only one time. After viewing the scene you will be given a memory test asking you to pick from a list of 10 words the only words appearing in the scene." The test was used to motivate the subject to look for the text in the scene and to motivate a search for the text more than an intensive text reading task, although the test results were discarded.

**Initial observation point and experiment end condition:** Subjects were all instructed to look at the scene with the same initial orientation, by looking straight to an initial point in the scene (similarly to how subjects are initially set by visualizing a target cross in the center of the screen in desktop eye tracking experiments). Regarding the experiment end condition, the experiment had no time limit allowing all subjects to fully look at the scene up to one complete 360 degree scene view. The rational was to allow all subjects to freely look at the scene each one at his own speed, up to look it fully but not look at the scene twice. Subjects were further reminded of the experiment end by voice commands.

**Figure 5.3:** A sequence of 12 selected frames taken by the Mobile Eye Tracker's scene camera. Following the frame sequence by rows, from the top left to the bottom right, a typical pattern of a subject making a full 360 degrees scan of the scene while searching for text objects. The FOA position is superimposed in red color.

**Calibration:** Eye tracker calibration was performed before running each experiment in a "calibration location" distinct from the experiment locations. Data was stored on a notebook computer which participants carried in a backpack during both calibration and experiment.

## 5.3 Data processing

Exploiting the experimental design described in Sect. 5.2, here we introduce a state-of-the-art procedure to map the raw data eye position from a per-frame reference system of the mobile eye tracker to a global reference system shared between all subjects.

### 5.3.1 Event detection

Event detection is performed using the software tools provided with the recording device. The built-in detector looks first for fixations, using a dispersion based algorithm to cluster row data points in a fixational average point. Other events are derived from them. A saccade is regarded as being bordered by two saccades, a blink is considered a special case of fixation where eye data is not present.

Data from mobile eye tracking are likely to have some missing sequence of data, due to the impossibility to track the focus of attention (FOA) position under certain illumination condition. We observed that both high intensity light saturating the ETG eye camera or clouds-filtered infra-red light is able to affect the eye recording process.

### 5.3.2 Mapping

We used a state-of-the-art procedure to compute homography transformation between each frame and the panoramic image, based on the detection of stable image key points. The eye tracker scene camera records high resolution videos at 24 fps, but among those we only needed to compute the homography at frames for which there have been detected fixation events. The homography matrix is then used to transform the FOA coordinates from the video frame coordinate system to the panoramic picture reference system.

We used the OpenCV library to compute the homography transformation and the key point detectors and descriptors. The high resolution panoramic picture lead to a large number of detected key points, however we have not made any reduction on the number of the detected key points as we are interested in maximise the correctness of homography transformations at the cost of a higher computational time. As the slowest part of the algorithm was found to be the computation of pairwise feature descriptor distance, represented in Fig. 5.4 by the blue lines connecting pairs of key points between the frame and the panoramic picture. To speed-up this computation we used a cosine approximation of the feature distance.

Results of the homography transformation have been checked by visual inspection. More than the 80% of the frames resulted correctly mapped. To ensure the full correctness of the mapping, we developed a tool for manual mapping of the FOA position. Incorrect homography transformations are mainly due to the presence of

shadows cast on the ground by trees and foliage, affecting a large part of the scene in the eye tracker's scene camera, and making it appear different from the panoramic scene. We observed that due to the large width of the scene view, small changes, such as the presence of people walking, did not affect the correct mapping.

The mapping process is summarized in Fig. 5.5. First is retrieved a video frame taken at the fixational event timestamp, then it's computed the homography transformation. Frame and the current FOA are mapped on the panoramic picture. Procedure is repeated iterating on all the fixational events and the final mapped scanpath is shown in Fig 5.5(d).

**Figure 5.4:** Graphical illustration of the key points matching procedure. Key points from a low resolution taken by the ETG device (top left figure) are matched to the key points in a high resolution panoramic picture (bottom figure). Lines represents pairs of key points scoring highest in the matching.

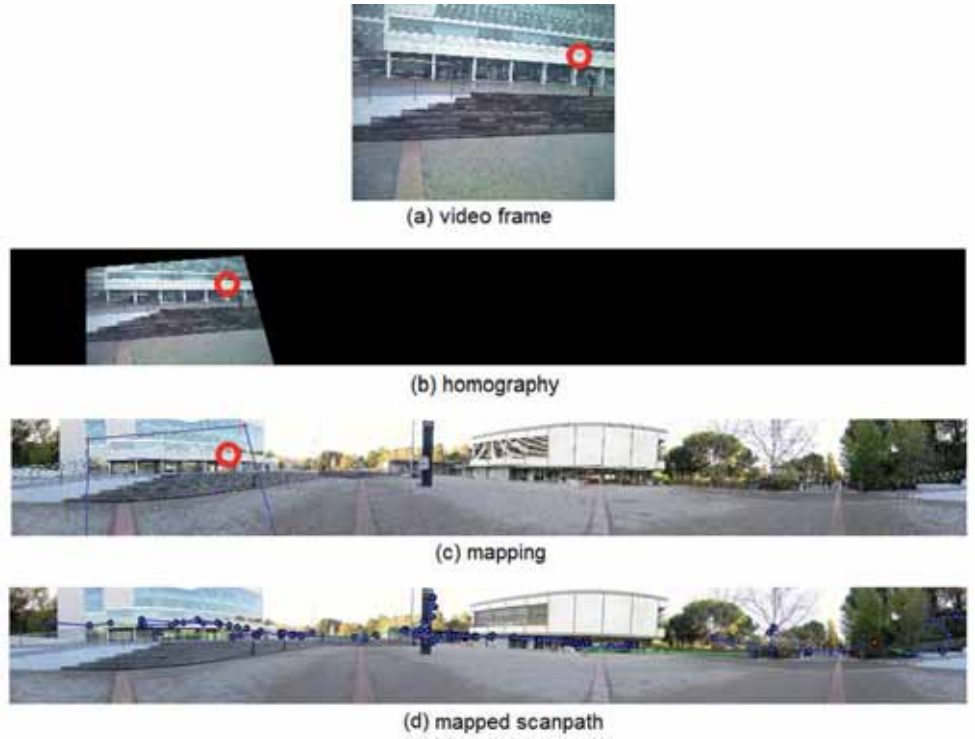



**Figure 5.5:** Homography transformation, the procedure to map a frame onto the scene. (a) a video frame taken at the event timestamp, (b) homography transformation video frame and foa position, (c) mapping of the FOA on the panoramic picture, in blue the video frame mapping for visualization purpose, (d) scanpath after mapping all fixation on the panoramic picture

## 5.4 Model simulations and Performance

In this section we assess the model performance and evaluate how changes to some parts of the model are affecting the global model performance. We specifically look at the effect of biasing the baseline model presented in Chapter 3 with different settings of 1) amplitude distributions as in sec 5.4.3, 2) angle distribution as in sec 5.4.4, 3) reward mechanism as in sec 5.4.5.

The amplitude and angle distributions are learned from the eye tracking data as obtained after the mapping described in section 5.3. Different sets of distributions lead to different sets of models. Aiming to investigate the effect of amplitude and angle biases, we look at differences in a per-location specific settings or in a global settings. In the first case using sets of distributions learned from a specific location and averaged over all the subjects, in the second case using sets of distributions learned from all the data averaging over all locations and all subjects. For easy of description the models are named with a three characters string coding in the form Model x-y-z, in which x,y,z stands for the amplitude, angle, reward parameter setting as further described in the appropriate following sections.

The scanpaths produced by the different sets of models are compared one with the other by employing a definition of performance accounting for how accurate are the fixations in hitting the in-scene text targets. Details on the performance metric are described in the Sect. 5.4.2. To account for the inherent stochasticity of eye movements, the human performance is computed as an average over 14 eye tracking experiments it's thought as the upper bound to the model performance. The chance level is computed as an average performance on 14 random sequences of fixational points and it's regarded here as the lower bound to model performance. Results are assessed individually per each location among the three locations of interest in this evaluation and depicted in Fig. 5.1.

### 5.4.1 Learning model parameters

Eye tracking data can be used to learn human oculomotor bias in the form of the amplitude and angle distributions of gaze shifts. As long as the model describes the gaze shift process as a two state finite state machine, switching between local exploration (short saccades) and large relocations (long saccades), both the amplitude and angle distributions need to be specified for the two states. We split the sequence of saccadic eye movements data in two sets of short and long saccades using an algorithm of robust clustering [2]. The distributions for amplitude and angle distribution are than computed twice on the two sets of data.

**Amplitude distribution:** According to the baseline model introduced in Chapter 3 the gaze shift amplitudes follows the $\alpha$-stable distribution and the $\alpha$-stable parameters can be fitted using the Chambers, Mallows, and Stuck procedure [22]. The stable fitting procedure returns an estimate of the four parameters in a fit of the $\alpha$-stable distribution to the amplitude of gaze shift eye tracking data. Parametrization is given in terms of characteristic exponent *alpha*, the skewness *beta*, the scale *gamma* and the location *delta*.

**Angle distribution:** The distribution of angles is modeled in the form of a

transition matrix describing the sequence of angle directions as a discrete-time Markov chain (DTMC) over a finite state space. Eye tracking data are used to build the transition probabilities from the state encoded by the angle at the time t-1 to the angle at time t. Angle directions are quantized in 16 directions. Sample transition matrix learned from all the data at the location 1 are visualized in Fig. 5.6.

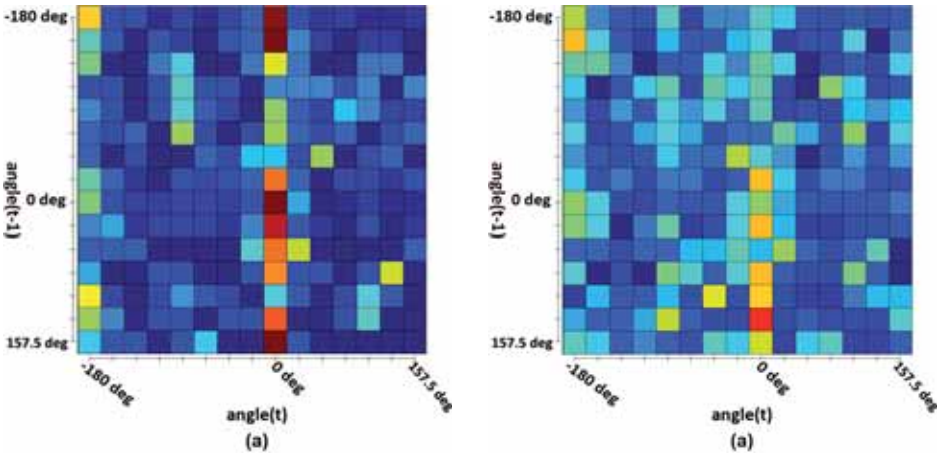



**Figure 5.6:** Histogram-based transition matrix of angle occurrence. (a) local exploration, (b) large relocations. Both transition matrix are computed on all the fixation data at the location 1. Rows represents angles at time t-1 and columns angles at time t. Values are normalized per-row to sum 1.

### 5.4.2 Performance metric

Performance is measured in terms of how accurate are the fixation in hitting text targets. The metric used here is an adaptation from the metric used by Judd [54, 55], although we applied it in a complementary way, as in her work the objective is to score saliency maps, as her proposed algorithm predicts the saliency map per pixel. We keep the idea that human fixations can be used to score the saliency maps as a function of the correlation between fixation locations the saliency map. In Judd the evaluation procedure consists in the following steps: The saliency map is thresholded at n =1, 3, 5, 10, 15, 20, 25, and 30 percent of the image for binary saliency maps which are typically relevant for applications. For each binary map, they find the percentage of human fixations within the salient areas of the map as the measure of performance. As the percentage of the image considered salient goes to 100%, the percentage of human fixations within the salient locations also goes to 100% [54, 55].

Our adaptation of this procedure accounts for two facts: 1. our algorithm generates a scanpath and does not a saliency map, 2. we are interested to assess the capability of the system to generate fixation related to text location. In such a sense the adaptation we propose consist in replacing the roles proposed in Judd [54, 55] of the saliency map with a distance transform and the human eye fixations recorded in eye tracking sessions with the model generated fixations.

The distance transform is an operator defined on a binary map, and consists in labeling each pixel of the binary map with the distance to the nearest obstacle pixel. A graphical representation of how we use the distance transform is provided in Fig. 5.7. We use as binary map a ground truth map, labelling text pixels in white and non-text pixels in black as shown in Fig. 5.7(a). The distance transform is computed as distance for each pixel to the nearest text pixel, giving as result a gray level map as in Fig. 5.7(b) in which the light gray levels encode distance close to zero, and dark gray levels encode bigger distance from text. The distance transform is then thresholded such that a percent of the image pixels are above the threshold. Similarly to Judd, such percentages are set to the $\{1, 3, 5, 10, 15, 20, 25, 30\}$ percent of the total number of image pixels. The thresholded maps corresponding to the levels of 1%, 5% and 15% are depicted in Fig. 5.7(c,d,e). The higher is the threshold value, the bigger is the selected portion of the distance transform (drown in white). The thresholded maps are used as binary classifier on every fixation, counting them as on-target or off-target.

This use of the distance transform implements what we define here as a *relaxation factor*. As the multiple thresholded maps have the property of being enlarged version of the ground truth map, it allows to account for fixational points that are slightly off-target with a variable level of relaxation. This is very useful as allows as to account for unknowns in the eye tracking, especially for the fovea region extension that is here affected by variable eye-target distance and the accuracy of the eye tracker instrument in FOAs position detection.

**Figure 5.7:** (a) The binary ground truth map corresponding to the panoramic picture in Fig 5.1. The white pixels annotate the text, (b) Distance transform gray levels encode distance value for each pixel to the nearest text pixel, (c,d,e) effects of the relaxation factor at the levels of respectively 1%, 5%, 15%

### 5.4.3 The effect of learning amplitude distribution

In our model of attention the amplitude distribution of gaze shift is governed by the $\alpha$-stable distribution. Varying the $\alpha$-stable parameters has an important impact of the selection of visual information as they are directly linked to the next fixation generation and the selection of visual patches.

Using the Chambers, Mallows, and Stuck procedure [22] we can learn the $\alpha$-stable parameters from eye tracking data and study the effect of varying the $\alpha$-stable parameters on the system performance in text search. The amplitude settings studied here are the following two: 1) a location-generic setting, in which are used data from all the subjects and all the locations 2) a locations-specific setting in which are used the data from all the subjects and only a specific location. Following the baseline model, we do not include any bias to the saccadic angle directions, and angles are sampled from the uniform distribution. Regarding the reward mechanism, in this first analysis we assign zero reward independently of the detectable value. For easy of description the models are named with a three characters string coding that are the amplitude, angle and reward setting. The model are coded as Model S-U-N and G-U-N standing respectively for a {location-Specific amplitude, Uniform angle, No reward} setting and the second for {location-Generic, Uniform angle, No reward}.

Plots in Fig. 5.4.3 show the average performance of system measured as the true positive rate against the relaxation factor. Plots are one per locations (over the three locations on interest). Plotted curves are averaged values of true positive rate on 14 model simulation trials , 14 human subjects eye tracking sequences and 14 random sequence for the chance level. We can observe the following things: Location specific amplitude parameters lead to better performance than generic parameters. At location 1 and location 2 the S-U-N model is outperforming the G-U-N for any level of the relaxation factor.

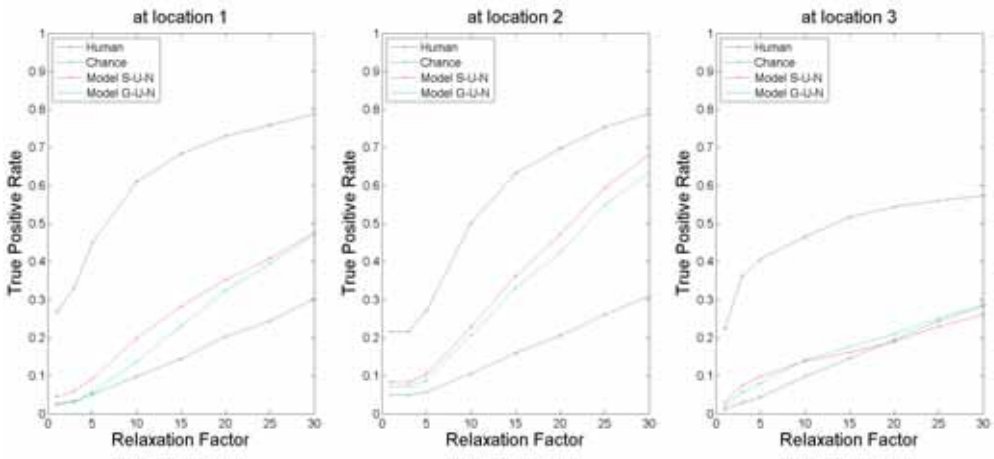



**Figure 5.8:** Average performance of the system, at the three different locations and under different sets of amplitude parameters

At location 3 the models S-U-N and G-U-N are very close each other although both are very close to the chance level. This might be explained in terms of text item density, as at the location 3 there are just a few text areas. In such a sense the third location is inherently more difficult than the other two, for the purpose of a text search task. This observation is also confirmed by looking at the human performance level that is close the level of 0.8 for both location 1 and 2, lower than 0.6 at location 3. In short learning amplitude parameter in a location-specific setting is a winning strategy, as we intuitively expected. We can read this as an indicator that amplitude distributions are, somehow, able to capture location specific distribution of objects as a form of scene layout.

### 5.4.4 The effect of learning angle distribution

In our model the angle distributions of gaze shift are modelled as uniform distribution in the range $(0, 2\pi)$ radian. As in the previous section, we study the effect of varying the angle distributions by learning from different sets of eye tracking data. The angle distribution settings are the following two: 1) a location-generic setting, in which are used the data from all the subjects and all the locations. 2) a locations-specific setting in which are used the data from all the subjects and only a specific location. Details on how we computed the angle distributions are described in Sect. 5.4.1. Regarding amplitude distribution we set location-specific amplitudes. Regarding the reward mechanism, here we always assign zero reward independently of the detected value. The models in this section are coded as the Model S-S-N and Model S-G-N standing for a {location-Specific amplitude, location-Specific angle, No reward} setting and the second is {location-Specific amplitude, location-generic angle, No reward}. Performance are reported in Fig. 5.4.4 and the Model S-U-N, already introduced in the previous section, is shown in the following for comparison purpose. We can observe the following things: The model S-S-N is always scoring higher than the S-G-N, showing that learning location-specific angle distribution is better than location-generic ones. In this respect we can say that location-specific bias are always better than location-generic ones, and this observation is true for both amplitudes and angles. In addition we can see that the uniform angle distribution is also scoring well at location 1 and 2, as shown in Fig. 5.4.4 S-U-N model. However at the location 3, only the S-S-N model is able to achieve a true positive rate significantly higher than the chance level. It seems that the location-specific angle distribution can hit text items more efficiently than the uniform distribution, especially in locations with very low text item density. In this respect it's difficult to give general conclusion, however an explanation might lie in strongly biasing toward horizontal saccadic direction observable in the location-specific angle distribution.

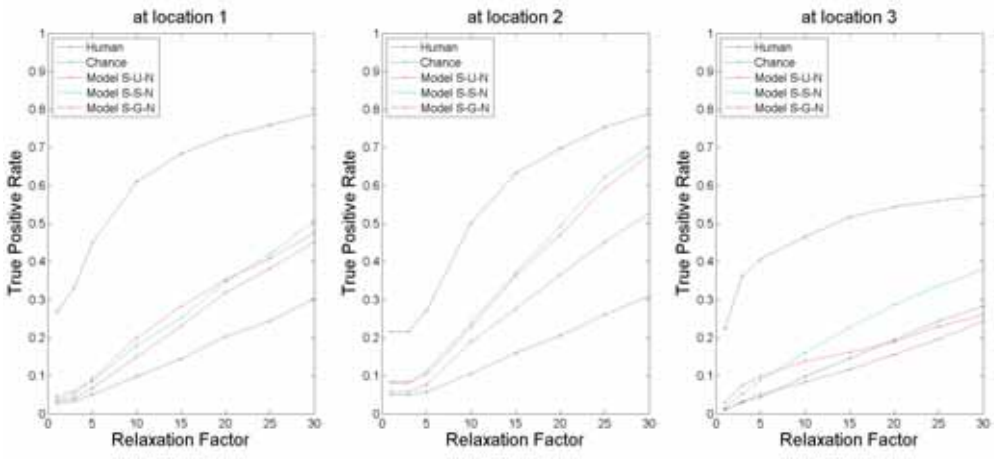



**Figure 5.9:** Average performance of the system, at the three different locations and under different sets of angle parameters

### 5.4.5 The effect of reward

Third component of this analysis is the role of the rewarding system. In an active vision paradigm be able to get feedback from the scene and choose where to look next, is of core importance as allowing to spend time on portions of the scene that are relevant in the sense of the task related value. Differently from previous analysis in which the attentive system was set to be agnostic to text presence, here we employed a state-of-the-art text detector [40, 56], and a perfect classifier (or oracle) always giving the correct answer on the text presence. Regarding amplitude and angle distributions we set location-specific amplitudes and location-specific angle. The models in this section are coded as the Model S-S-A and Model S-G-P in which the first stands for {location-Specific amplitude, location-Specific angle, state-of-the-art text detector based reward} setting, the second for {location- Specific amplitude, location-Specific angle, perfect classifier based reward}. Performance are reported in Fig. 5.4.5. The S-S-A model, based on a state-of-the-art text detector, is able slightly improving over the S-S-N. Although small this is a positive result as a correct interpretation of results must take into account the discriminability of the text in real scene images. Although the text detection method used here is scoring as top performance (precision = 0.58, recall = 0.54 on the MSRA-TD500 dataset [40]), the scene illumination conditions are strongly affecting the text appearance. The S-G-P model, based on a perfect classifier, is instead outperforming any other model, although never reaching the true positive rate achieved by the human level. This is something expected as the role of the perfect classifier is to inform the attention model on a per-fixation basis, although its role is not to ensure the selection of on-target fixation. Finally the human performance is always higher than any achievable performance by our models, as accounting contextual information by far larger then the one included in our models.

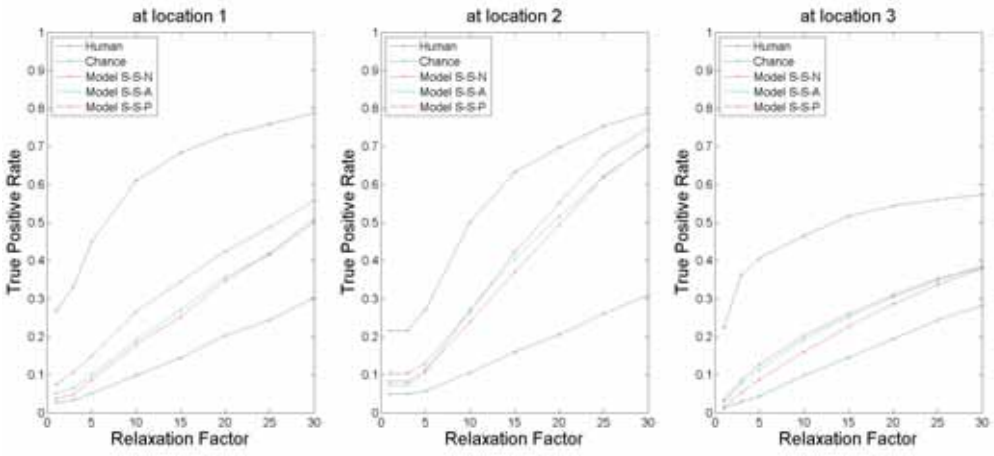



**Figure 5.10:** Average performance of the system, at the three different locations. Accounting for the effect of reward

### 5.4.6 Comparison

Here we report a comparison of the different models introduced in this chapter. The names of the models are coded as described in the previous sections. Plots in figure 5.11 report a very dense picture of the system performance averaged over the three locations. We observe the following: 1) As long as the human performance is around 0.7 for a relaxation factor ranging in (20, 30)%, we register an average upper bound to performance of around seven fixations over ten for almost on target fixations. 2) The use of location-specific biases is improving over the use location-generic biases. The S-U-N model scores better than the G-U-N for any relaxation factor (using as using location-specific amplitudes, and uniform angle distribution). 3) The S-S-N model is improving over the previous two as including both location-specific amplitudes and angles. The models S-U-N and G-U-N, both characterized by an uniform angle distributions, score at an intermediate level. Although not performing equally well for all locations as observed in previous section. 4) Learning location-specific angle distributions (as in the model S-S-N) improves performance, especially at higher levels of relaxation factor indicating the tendency of having slightly off-target fixation. 5) The use of a rewarding mechanism is improving performance if reliable. The use of a rewarding system based on a state-of-the-art classifier makes a small although positive improvement over not using any reward.

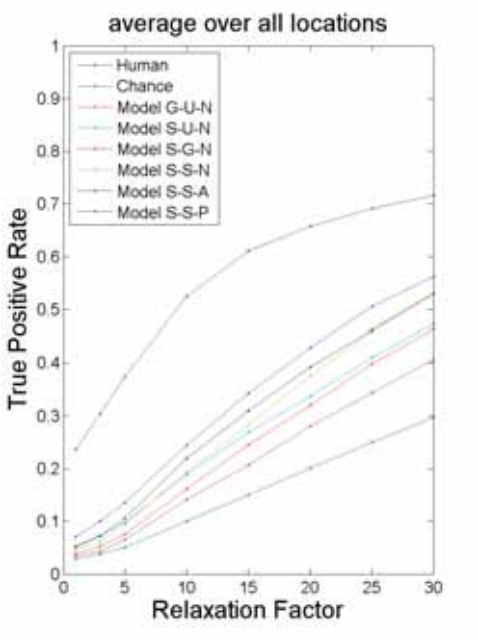



**Figure 5.11:** Average performance over all subjects and all locations. Comparison of performances for variations to the baseline models. Humans and chance performance level are plot for comparison purpose.

## 5.5 Conclusion

In this Chapter we have studied to what extent our model of attention can account for human eye movements in a close to real life exploration of a 360 degree scene. We have specifically addressed the condition of highly incomplete visual information, fully embracing an active vision paradigm. We stressed the point of limited visual information and deeper investigate the mechanism through which perceptual bias can modulate object search performance.

We use a mobile eye tracker to record eye data from human subject in an outdoor setting. In such experimental condition the scene is not all visible at a glance and head movements as well body rotation are needed to fully see the scene in its 360 degrees. We developed a novel and ad-hoc methodology to compare model simulated eye data with the human eye data from mobile eye tracking recordings. A novel element in our analysis is in having resorted to a procedure based on the computation of homography transformation to map the mobile device camera view, to a panoramic picture. This operation allowed us to think in bi-dimensional reference system, and establish a clear correspondence between mobile eye tracking data and model generated data.

Experiments show that learning biases in a location-specific settings translates in higher average performance. Location specific amplitude parameters lead to better performance than generic parameters, the inclusion of location-specific angle distributions seems to be able to achieve higher true positive rate by properly biasing the saccadic direction in locations with very low text item density. The use of a reliable rewarding system allows to select significantly better the informative regions of the scene, leading to the closest to human performance, although the latter is accounting contextual information by far larger then the one included in our model.

# Chapter 6

# Conclusions

In this thesis we have presented an integrated computational model of eye guidance for task-dependent attention deployment to objects in natural pictures. To the best of our knowledge, the model is novel in proposing a unified framework that i) accounts for task-dependent visual attention on semantically rich natural images by using different levels of representation, beyond the baseline saliency maps; ii) simulates gaze shifts that exhibit statistical properties close to those of eye-tracked subjects, by extending previous approaches proposed in the literature, [10, 11] that addressed the intrinsic stochasticity of gaze shifts; iii) tackles close to real life experimental condition characterized by high deception and incomplete information scenarios:

- **task-dependent visual attention:** For what concerns the task-dependent visual attention, the proposed model can cope with eye guidance both under a search task, an issue which has been taken into account by some models [128, 85, 71, 133], and under a generic picture viewing task, which has been typically accounted for by saliency/relevance-based models (either bottom-up [53] or top-down biased [108, 24]). The key to such integration is that, different from those models, we have considered the generation of a scan path as the interplay among several levels of representation and control that goes beyond the classic debate bottom-up vs. top-down, but brings payoff, value and motor representations into the game. We believe that, although this broad and flexible approach also creates new theoretical and computational challenges, this very breadth is an important issue to address. In fact, to succeed in complex environments we must act in a flexible manner as appropriate for a given task, which suggests that a stage of visual selection can be distinct from that of saccade motor selection. For instance, the priority map may encode signals of visual selection that are not eventually captured by current action based decision module. Differently from methods using purely visual top-down modules, the biases provided do not amount directly to motor command, and action related areas may also block or supplement its signal as required in a given task. This integration of different levels of representation and control is important to define some issues that remain elusive if considered with respect to a single level or locus. One such example is the inhibition of return (IOR). Depending on the task, different

variants of IOR exist [126]. In our model, the priority map explicitly uses IOR in a classic way, by suppressing the response at the currently attended location. However, a reduction or an enhancement in the reward likelihood can modulate the IOR at the priority map level. In general, this multiple level interaction can be a way of framing the discussion surrounding a functional interpretation of IOR (that is fostering or not optimal foraging behaviour, see [126]).

More generally, the use of value and payoff can provide a suitable bridge to explain gaze behaviour that, even in the absence of given task seems to be driven by internal *motivational saliency*, which in pathological conditions could be generated by a disruption of biological reward systems [18]. Visual scan path analyses provide important information about attention allocation and attention shifting during visual exploration of social situations characterised by both cognitive complexity and emotional content or even strain. On the one hand, the approach proposed here paves the way to the effective exploitation of computational attention models in the emerging domain of social signal processing [119], and, more broadly, to cope with the problem the affective modulation of the visual processing stream [79, 80] with the aim of closing the gap between emotion and cognition [42]

- **scan path variability:** As regards the scan path variability, the model attempts at filling a gap in the current computational literature (cfr., [12]). The majority of models in computational vision basically resort to deterministic mechanisms to realise gaze shifts, therefore, if the same saliency map is provided as input, they will basically generate the same scan path. Moreover ignoring motor strategies and tendencies that characterise gaze shift programming results in distributions of gaze shift amplitudes different from those that can be observed from eye-tracking experiments. We have presented in examples showing that the overall distributions of human and model generated shifts are close in their statistics.

  The core of such strategy actually relies upon a mixture of $\alpha$-stable motions modulated by the different visuomotor levels of control participating to the action-perception loop. The composition of random walks in terms of a mixture of $\alpha$-stable components allows to treat different types of eyes movement within the same framework and makes a step towards the unified modelling of different kinds of gaze shifts. The latter is a research trend that is recently gaining currency in the eye movement realm [78]. For instance, when Eq. (3.16) is exploited for within-patch exploration, it generates a first-order Markov process, which is compatible with most recent findings [6]. Notice that this approach may be exploited for a principled modelling of individual differences and departure from optimality [70] since providing cues for defining the informal notion of scan path idiosyncrasy in terms of individual gaze shift distribution parameters. The latter represents a crucial issue both for theory [97, 112] and applications [65]. For instance, the study by Sprenger *et al.* [105], concerning patients with schizophrenia, has shown that show that alterations such as restricted free visual exploration were present in patients independently of cognitive complexity, emotional strain or physical properties of visual cues implying that they rep-

resent a rather general deficit, which may be accounted for in terms of group specific oculomotor bias or scanning strategy.

- **Close to real life experimental conditions:** From a searcher standpoint real life experimental conditions are typically characterized by high deception and incomplete visual information. As long as target in real scene are typically not easy to distinguish due to the huge variability in object appearance, relocations of the perception point of view is a straightforward strategy to gather multiple views of the target and, hopefully, more easily detect the object. As soon as adopting an active vision paradigm, the limited visual information has to be accounted as an intrinsic property of a vision system implementing visual information selection by interaction with the scene. Investigating in this direction we collected a dataset of human eye movements in mobile eye tracking experiments in outdoor settings. Our attentive system allowed to simulate the limited visual information and the gathering of small scene views from a large panoramic picture. Performance, measured in terms of true positive rate of text targets, report the human upper bound performance at the level of about 7 fixations on-target each 10 fixations (0.7 %). Simulations show that the model performance are largely lower than human performance as measured through eye tracking data. Model performance can be improved by accounting for oculomotor bias in amplitudes and angle distributions of gaze shifts achieving a true positive rate in the range (0.4; 0.5) for a relaxation factor in range (20, 30) %. The use of a reliable rewarding system allows to reach a bit higher performance of about still far from the human upper bound. Such result is not surprising as the visual system is able to account contextual information by far larger then the one included in our model.

## 6.1 Limitations and Future Perspective

Clearly, there are some limitations of the model in its present version. We do not consider here time-varying or multiple task assignment, which may be important in real world behaviours. Also, we barely touched the level of neural implementation. However, in this respect, the model is agnostic about whether or not probabilistic computations can be neurally implemented (see the review by Knill and Pouget [58]). This is an intriguing but intricate debate. For instance, Heinke and Humphreys [44] raised the interesting point of using differential equations the exhibits chaotic behaviour to account for noise and recently Churchland and Abbot [26] argued that randomness in neuronal firing rates and spike timing could arise from a network built of deterministic neurons with balanced excitation and inhibition. Further, to make the broad integration behind the model feasible, we have focused on the core issues, providing some black-box or simulated implementations for other components. For instance, for the text localisation/detection task we rely on simulated detectors both for the pre-attentive coarse grained localisation and for the fine-grained detection/recognition. In a preliminary work using a simpler version of the model presented here [27] we have experimented with a text localiser component based on a Relevance Vector Machine classifier applied to "gist" texture features *á la* Torralba [114] both at a coarse and at

a high resolution level. However, "textual objects" are a difficult task as opposed to faces for which, at least, efficient and effective face detectors do exist [120], if one is not concerned with the biological plausibility of the algorithm. Actually, our current research work is indeed addressed at verifying the suitability of our model in a difficult practical problem such as text localisation and detection "in the wild", in order to overcome present limitations of attentive-based approaches proposed within such realm [27]. To this end, we are adapting the model to handle time-varying images, and we are performing mobile eye-tracking experiments outside the lab, in complex urban environment. Another limitation, which is conceptually more important than the previous one, is that using value and payoff calls for adopting learning procedures that could be at hand with such information and could be exploited, in the case of a search task, for priming the guidance process [104]. However, it is clear that when dealing with restricted real-world tasks (e.g., crossing a road or making a tea cup) the learning stage can be effectively stated; what has to be learned in the task of searching in a dataset of mostly unrelated pictures of natural scenes is less evident. Treatment of these topics is deferred to a future study.

We do not by any means regard the following as a complete picture of what actually goes on in the attentive brain. But results presented here encourage us to put forth this preliminary attempt at outlining a theoretical foundation grounded in a principled integration of several levels of representation and control for supporting eye guidance, albeit calling for further research into these basic processes.

# Appendix A

# Foraging models and Lévy flights

## A.1 The foraging metaphor

The foraging metaphor comes from the assumption that animals are in some way optimizing in their foraging activities [23]. Several authors have designed mathematical models to predict the foraging behavior of animals and they all assume that the fitness of the forager is a function of the efficiency of the foraging and that natural selection has resulted in animals that forage so as to maximize this fitness [84]. These models have become known as "optimal foraging models" and are traditionally divided in four categories: optimal diet, optimal patch choice, optimal allocation of time to patches, optimal patterns.

### A.1.1 Primates foraging model

Recently Boyer and others [13] introduced a simple foraging model for individual primates foraging in forest. The foraging environment is modeled as a two-dimensional square domain containing N point-like targets randomly and independently distributed in space (Poisson process), representing the trees with fruits that monkeys eat. Each target is identified by an index $i$ and the target's size (or fruit content) is described by $k_i$ which one, according to recent work (Enquist and Niklas 2001; Niklas et al. 2003), is assumed to be distributed according to a inverse power-law probability distribution

$$P(k) = CK^{-\beta} \tag{A.1}$$

where $C = 1/\sum_{k=1}^{\inf} k^{-\beta}$ is the normalization factor and $1 < \beta < inf$ is a fixed exponent characterizing the environment.

A forager located at a starting point (the centre of the domain), knows the location and size of all targets in the system, and recursively follows some rules of motion:

(a) the forager located at the target number i will move in a straight line to a target j such that the quantity $l_{ij}/k_j$ is minimal among all available targets $j \neq i$ in the system, where $l_{ij}$ is the distance separating the two targets and $k_j$ is the size of target $j$;

(b) the forager does not choose an already visited target, as it is assumed that visited targets no longer contain fruits.

#### A.1.1.1 model's assumption

1. The distance size ratio $l/k$ roughly represents a cost/gain ratio for a move motivating the being attracted or repelled from a certain site, so that valuable targets may be chosen even if they are not the nearest to the monkey's position.

2. A crucial assumption of this model is that the forager is assumed to know location and size of all targets to move from one target to another. Although, according to previous work, many animals (bees, rodents, primates) do not forage randomly but rely instead on cognitive maps to navigate their environment (Collett et al. 1986; Garber 1989; Dyer 1994). These maps may contain information on the location of different targets and the geometric relationships between them (Kamil & Jones 1997).

   In order to relax the assumption of perfect knowledge by the foragers, we assume that the forager is inexact in evaluating the distance to a given target, as well as its fruit content. Models simulation concludes that when the foraging rule is imperfect and with a typical error of 65%, no qualitative differences in the distributions are found with respect to the perfect, deterministic rule.

#### A.1.1.2 findings

By varying its main parameter $\beta$ which describes the decay of the tree-size frequency distribution, the trajectories of the artificial forager following a differs widely. Levy walks arise as a consequence of food intake maximization in a spatially disordered, heterogeneous environment where the location of resources is at least partially known.

## A.2 Efficient search in foraging

### A.2.1 Random walks and Lévy flights in a nutshell

In continuous time a $d$-dimensional random motion under the influence of a force field for the stochastic variable $\mathbf{r}(t)$ can be described by the Ito-type Stochastic Differential Equation (SDE) [38]

$$\mathbf{dr}(t) = \mathbf{g}(\mathbf{r}, t)dt + \mathbf{D}(\mathbf{r}, t)\xi dt. \tag{A.2}$$

The trajectory of the variable $\mathbf{r}$ is determined by a deterministic part $\mathbf{g}$, the drift, and a stochastic part $\mathbf{D}(\mathbf{r}, t)\xi dt$, where $\xi$ is a random vector and $\mathbf{D}$ is a weighting factor. Thus, clearly, the motion is determined by the probability density function $f$ from which $\xi$ is sampled. For instance, if $f$ is a Gaussian distribution, the usual Brownian motion occurs.

However, Brownian motion is nothing but a special case within the family of stochastic processes qualifying as natural models for random noise sources. Other types of motion can be generated by resorting to the class of the so called $\alpha$-stable

distributions [39]. These form a four-parameter family of continuous probability densities, say $f(\xi;\alpha,\beta,\gamma,\delta)$, parametrized by the skewness $\beta$ (measure of asymmetry), scale $\gamma$ (width of the distribution) and location parameters $\delta$ and, most important, the characteristic exponent $\alpha$ or index of the distribution that specifies the asymptotic behavior of the distribution.

Indeed the probability density function $p(l)$ of length jumps scales, asymptotically, as $l^{-1-\alpha}$, and thus relatively long jumps are more likely when $\alpha$ is small. By sampling $\xi \sim f(\xi;\alpha,\beta,\gamma,\delta)$, for $\alpha \geq 2$ the usual random walk (Brownian motion) occurs; if $\alpha < 2$ , the distribution of length jump is "broad" and the so called Levý flights take place.

More precisely, a random variable $X$ is said to have a stable distribution if the parameters of its probability density function (pdf) $f(x;\alpha,\beta,\gamma,\delta)$ are in the following ranges $\alpha \in (0;2]$, $\beta \in [-1;1]$, $\gamma > 0$, $\delta \in \mathbb{R}$ and if its characteristic function $E\left[\exp(itx)\right] = \int_{\mathbb{R}} \exp(itx)dF(x)$, $F$ being the cumulative distribution function (CDF), can be written as

$$E\left[\exp(itx)\right] = \begin{cases} \exp(-|\gamma t|^{\alpha})(1 - i\beta\frac{t}{|t|})\tan(\frac{\pi\alpha}{2}) + i\delta t) \\ \exp(-|\gamma t|(1 + i\beta\frac{2}{\pi}\frac{t}{|t|}\ln|t|) + i\delta t) \end{cases}$$

the first expression holding if $\alpha \neq 1$, the second if $\alpha = 1$.

Special cases of stable distributions whose pdf can be written analytically, are given for $\alpha = 2$, the normal distribution $f(x;2,0,\gamma,\delta)$, for $\alpha = 1$, the Cauchy distribution $f(x;1,0,\gamma,\delta)$, and for $\alpha = 0.5$, the Lévy distribution $f(x;0.5,1,\gamma,\delta)$; for all other cases, only the characteristic function is available in closed form, and numerical approximation techniques must be adopted for both sampling and parameter estimation [22, 74, 62].

Special cases of stable distributions, for which the pdf can be written analytically, are the normal distribution ($\alpha = 2$), the Cauchy distribution ($\alpha = 1$), and the Lévy distribution ($\alpha = 0.5$) ; for all other cases, only the characteristic function is available in closed form, and numerical approximation techniques must be adopted for sampling and parameter estimation [22, 125, 74, 62].

Examples of Lévy flights, typically exhibiting local walk interleaved with long jumps, are presented in Figure A.1 (second and third plots) compared to Brownian motion (top left). In the same figure, the bottom right plot illustrates a random walk pattern obtained as a composite process simulated by sampling from a mixture of two $\alpha$-stable distributions indexed by $\alpha_1 = 2$ and $\alpha_2 = 1$, respectively, and mixture weights $w_1 = 0.4, w_2 = 0.6$. It is worth noting in the latter case that the walking pattern could be identified as a Levy pattern though, in contrast with the other cases, the pattern generating process is not a pure Levy process, but a composite one (Brownian and Cauchy).

Coming back to (A.2), in many applications [9] $\mathbf{g}(\mathbf{r},t)$ is modelled as a force field due to a potential $V(\mathbf{r},t)$, that is $\mathbf{g}(\mathbf{r},t) = -\nabla V(\mathbf{r},t)$, Then (A.2) can be written as a Langevin equation,

$$\mathbf{dr}(t) = -\nabla V(\mathbf{r},t)dt + \mathbf{D}(\mathbf{r},t)\xi dt \tag{A.3}$$

In this case the probability density function $P(\mathbf{r},t)$ can be obtained via differential equations of the Fokker-Planck type [95] for Brownian motion and via their fractional versions in case of Lévy flights [68].

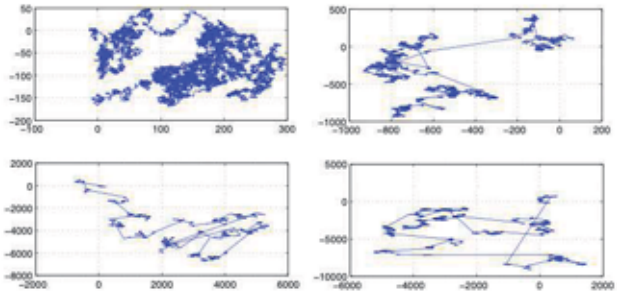


**Figure A.1:** Different random walks (left column) obtained by sampling $\xi_\alpha$ for different $\alpha$ parameters; the walks shown in the top left,top right and bottom left plots have been generated via $\alpha = 2$, $\alpha = 1.6$ , $\alpha = 1$, respectively; the bottom right plot, represents a composite walk sampled from a mixture of two stable distributions indexed by $\alpha = 2$ and $\alpha = 1$, parameters.

Stochastic differential equations have been used in researches concerning memory retrieval, language comprehension, and visual search (see, e.g., [87], [137] ), and the concept of potential functions is used in ecological modelling of animal movements to motivate a form for the drift term as a function of distances to selected habitat covariates [83]. In particular Lévy flights have been used to models searches of foraging animals and they have been shown to produce optimal searches in term of the ratio between the number of sites visited to the total distance traversed by a forager [100], [121].

### A.2.2 Efficiency

In a seminal paper [14], Brockmann and Geisel have shown that a visual system producing Lévy flights implements a more efficient strategy of shifting gaze in a random visual environment than any strategy employing a typical scale in gaze-shift magnitudes; evidence of Lévy diffusive behavior of scanpath has been presented in [106]. Equation (A.3) has been used in [9], within a "foraging metaphor", assuming that the Lévy-like property of scanpath mirrors patterns of foraging behavior found in many animal species [121]: the stochastic part of the motion was generated by Cauchy noise and the potential was designed as a function of a saliency map to derive a gaze-shift model (see [9] for a detailed discussion, and [66],[69] for application to robot vision relying on stochastic attention selection mechanisms).

However, the general applicability of Lévy flights in ecology and biological sciences is still open to debate, as recent experimental data show that the movement patterns of various marine predators and terrestrial animal exhibit a Lévy walk pattern in areas with low abundance of preys or foods and Brownian walk pattern (a sort of food tracking) in areas with high abundance. [28].

Thus, in complex environments, optimal searches should result from a mixed/composite strategy (generating patterns similar to the bottom right one of Fig. A.1), in which Brownian and Levy motions can be adopted depending on the structure of the land-

scape in which the organism moves [82], [93], [92]. Such strategy is optimal because Lévy flights are best suited for the location of randomly, sparsely distributed patches that once visited are depleted and Brownian motion gives the best results for the location of densely but random distributed within-patch resources [92].

# References


[1] C. Araujo, E. Kowler, and M. Pavel. Eye movements during visual search: The costs of choosing the optimal path. *Vision research*, 41(25-26):3613–3625, 2001. [Page 15]

[2] C. Archambeau and M. Verleysen. Robust bayesian clustering. *Neural Networks*, 20(1):129–138, 2007. [Page 68]

[3] Y. Wexler B. Epshtein, E. Ofek. Detecting text in natural scenes with stroke width transform. In *2010 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2963–2970, 2010. [Page 48]

[4] D.H. Ballard and M.M. Hayhoe. Modelling the role of task in the control of gaze. *Visual cognition*, 17(6-7):1185–1204, 2009. [Page 22]

[5] D.H. Ballard, M.M. Hayhoe, F. Li, S.D. Whitehead, JP Frisby, JG Taylor, and RB Fisher. Hand-eye coordination during sequential tasks [and discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 337(1281):331–339, 1992. [Page 23]

[6] M. Bettenbuhl, M. Rusconi, R. Engbert, and M. Holschneider. Bayesian selection of markov models for symbol sequences: Application to microsaccadic eye movements. *PLoS ONE*, 7(9):e43388, 2012. [Page 78]

[7] G. Boccignone. Nonparametric bayesian attentive video analysis. In *Proc. 19th International Conference on Pattern Recognition, ICPR 2008*, pages 1–4. IEEE Press, 2008. [Page 34]

[8] G. Boccignone, P. Campadelli, A. Ferrari, and G. Lipori. Boosted tracking in video. *Signal Processing Letters, IEEE*, 17(2):129–132, 2010. [Page 39]

[9] G. Boccignone and M. Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331(1-2):207–218, 2004. [Pages 83 and 84]

[10] Giuseppe Boccignone and Mario Ferraro. Ecological sampling of gaze shifts. *IEEE Trans. Systems Man Cybernetics - B*, pages 1–1, 2013. [Pages 8, 32, 34, 38, 39, 42, 43, 47, 48 and 77]

[11] Giuseppe Boccignone and Mario Ferraro. Feed and fly control of visual scanpaths for foveation image processing. *annals of telecommunications-annales des télécommunications*, 68(3-4):201–217, 2013. [Pages **37** and **77**]

[12] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013. [Pages **19** and **78**]

[13] D. Boyer, G. Ramos-Fernández, O. Miramontes, J.L. Mateos, G. Cocho, H. Larralde, H. Ramos, and F. Rojas. Scale-free foraging by primates emerges from their interaction with a complex environment. *Proceedings of the Royal Society B: Biological Sciences*, 273(1595):1743–1750, 2006. [Page **81**]

[14] D. Brockmann and T. Geisel. The ecology of gaze shifts. *Neurocomputing*, 32(1):643–650, 2000. [Page **84**]

[15] N. Bruce and J. Tsotsos. Saliency based on information maximization. *Advances in neural information processing systems*, 18:155, 2006. [Page **20**]

[16] S.S. Bucak, R. Jin, and AK. Jain. Multiple kernel learning for visual object recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1354–1369, July 2014. [Page **7**]

[17] M.S. Castelhano and K. Rayner. Eye movements during reading, visual search, scene perception: An overview. *Cognitive and cultural influences on eye movements*, pages 175–195, 2008. [Page **15**]

[18] Emily H Castellanos, Evonne Charboneau, Mary S Dietrich, Sohee Park, Brendan P Bradley, Karin Mogg, and Ronald L Cowan. Obese adults have visual attention bias for food cue images: evidence for altered reward system function. *International Journal of Obesity*, 33(9):1063–1073, 2009. [Page **78**]

[19] M. Cerf, E.P. Frady, and C. Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, 9(12), 2009. [Pages **22**, **25**, **32**, **36**, **39**, **47** and **52**]

[20] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. *Advances in neural information processing systems*, 20, 2008. [Pages **22** and **25**]

[21] M. Cerf, J. Harel, A. Huth, W. Einhäuser, and C. Koch. Decoding what people see from where they look: Predicting visual stimuli from scanpaths. *Attention in Cognitive Systems*, pages 15–26, 2009. [Pages **22** and **25**]

[22] JM Chambers, CL Mallows, and BW Stuck. A method for simulating stable random variables. *J. Am. Stat. Ass.*, 71(354):340–344, 1976. [Pages **43**, **68**, **71** and **83**]

[23] E. L. Charnov. Optimal foraging: the marginal value theorem. *Theoretical Population Biology*, 9:129–136, 1976. [Page **81**]

[24] D. A. Chernyak and L. W. Stark. Top–down guided eye movements. *IEEE Trans. Systems Man Cybernetics - B*, 31:514–522, 2001. [Page **77**]

[25] S. Chikkerur, T. Serre, C. Tan, and T. Poggio. What and where: A bayesian inference theory of attention. *Vision research*, 50(22):2233–2247, 2010. [Pages **22**, **32**, **33** and **39**]

[26] Mark M Churchland and LF Abbott. Two layers of neural variability. *Nature neuroscience*, 15(11):1472–1474, 2012. [Page **79**]

[27] Antonio Clavelli, Dimosthenis Karatzas, Josep Llados, Mario Ferraro, and Giuseppe Boccignone. Towards modelling an attention-based text localization process. In Joao. Sanches, Luisa Micó, and JaimeS. Cardoso, editors, *Pattern Recognition and Image Analysis*, volume 7887 of *Lecture Notes in Computer Science*, pages 296–303. Springer Berlin Heidelberg, 2013. [Pages **34**, **39**, **79** and **80**]

[28] E.A. Codling, M.J. Plank, and S. Benhamou. Random walk models in biology. *Journal of the Royal Society Interface*, 5(25):813, 2008. [Page **84**]

[29] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995. [Page **21**]

[30] J. Drewes, J. Trommershäuser, and K.R. Gegenfurtner. Parallel visual search and rapid animal detection in natural scenes. *Journal of Vision*, 11(2), 2011. [Page **25**]

[31] Miguel P Eckstein. Visual search: A retrospective. *Journal of Vision*, 11(5), 2011. [Page **8**]

[32] K.A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 17(6-7):945–978, 2009. [Page **21**]

[33] Wolfgang Einhäuser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14), 2008. [Pages **22** and **25**]

[34] Brian S Everitt. *The analysis of contingency tables*, volume 45. CRC Press, 2nd edition, 1992. [Page **57**]

[35] Tom Foulsham, Robert Teszka, and Alan Kingstone. Saccade control in natural images is shaped by the information visible at fixation: evidence from asymmetric gaze-contingent windows. *Attention, Perception, & Psychophysics*, 73(1):266–283, 2011. [Page **36**]

[36] S. Frintrop, E. Rome, and H.I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. on Applied Perception*, 7(1):6, 2010. [Page **22**]

[37] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of vision*, 8(7), 2008. [Page 21]

[38] C.W. Gardiner. *Handbook of stochastic methods*. Springer–Verlag, Berlin, Germany, 2011. [Page 82]

[39] B.V. Gnedenko and A.N. Kolmogórov. *Limit distributions for sums of independent random variables*. Addison-Wesley Pub. Co., 1954. [Page 83]

[40] Lluis Gomez and Dimosthenis Karatzas. Multi-script text extraction from natural scenes. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 467–471. IEEE, 2013. [Page 74]

[41] PM Greenwood and Raja Parasuraman. Scale of attentional focus in visual search. *Perception & Psychophysics*, 61(5):837–859, 1999. [Page 54]

[42] Claudius Gros. Cognition and emotion: perspectives of a closing gap. *Cognitive Computation*, 2(2):78–85, 2010. [Page 78]

[43] M.M. Hayhoe, A. Shrivastava, R. Mruczek, and J.B. Pelz. Visual memory and motor planning in a natural task. *Journal of Vision*, 3(1), 2003. [Page 22]

[44] Dietmar Heinke and Andreas Backhaus. Modelling visual search with the selective attention for identification model (vs-saim): a novel explanation for visual search asymmetries. *Cognitive computation*, 3(1):185–205, 2011. [Page 79]

[45] J.M. Henderson. Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11):498–504, 2003. [Page 23]

[46] J.M. Henderson, A. Pollatsek, and K. Rayner. Covert visual attention and extrafoveal information use during object identification. *Attention, Perception, & Psychophysics*, 45(3):196–208, 1989. [Page 15]

[47] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer. *Eye tracking: a comprehensive guide to methods and measures*. Oxford University Press, Oxford, UK, 2011. [Pages 37 and 54]

[48] T.S. Horowitz and J.M. Wolfe. Visual search has no memory. *Nature*, 394(6693):575–577, 1998. [Page 30]

[49] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Proceedings CVPR '07*, volume 1, pages 1–8, 2007. [Page 34]

[50] J.N. Ingram, K.P. Körding, I.S. Howard, and D.M. Wolpert. The statistics of natural hand movements. *Experimental brain research*, 188(2):223–236, 2008. [Page 26]

[51] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Advances in neural information processing systems*, 18:547, 2006. [Page 20]

[52] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews - Neuroscience*, 2:1–11, 2001. [Pages **20** and **22**]

[53] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1254–1259, 1998. [Pages **20**, **22**, **48** and **77**]

[54] Tilke Judd. *Understanding and predicting where people look in images*. PhD thesis, Massachusetts Institute of Technology, 2011. [Page **69**]

[55] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009. [Page **69**]

[56] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, Lluis Pere de las Heras, et al. Icdar 2013 robust reading competition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1484–1493. IEEE, 2013. [Page **74**]

[57] W. Kienzle, F. Wichmann, B. Schölkopf, and M. Franz. A nonparametric approach to bottom-up visual saliency. In *In Proc. NIPS 19*. MIT Press, 2007. [Page **20**]

[58] David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719, 2004. [Page **79**]

[59] D.C. Knill, D. Kersten, and A. Yuille. Introduction: A bayesian formulation of visual perception. In D.C Knill and W. Richards, editors, *Perception as Bayesian Inference*, pages 1–21. Cambridge University Press, 1996. [Page **30**]

[60] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4(4):219–27, 1985. [Page **20**]

[61] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, Cambridge, MA, 2009. [Pages **30** and **36**]

[62] I.A. Koutrouvelis. Regression-type estimation of the parameters of stable laws. *Journal of the American Statistical Association*, pages 918–928, 1980. [Page **83**]

[63] E. Kowler. Eye movements: The past 25 years. *Vision Research*, 51(13):1457–1483, 2011. 50th Anniversary Special Issue of Vision Research - Volume 2. [Pages **13** and **14**]

[64] A. Krause and C. Guestrin. Optimal value of information in graphical models. *Journal of Artificial Intelligence Research*, 35:557–591, 2009. [Page **35**]

[65] O. Le Meur, T. Baccino, and A. Roumy. Prediction of the inter-observer visual congruency (iovc) and application to image ranking. In *Proc. 19th ACM international conference on Multimedia*, pages 373–382, 2011. [Page **78**]

[66] H. Martinez, M. Lungarella, and R. Pfeifer. Stochastic Extension to the Attention-Selection System for the iCub. *University of Zurich, Tech. Rep*, 2008. [Page 84]

[67] E. Matin. Saccadic suppression: a review and an analysis. *Psychological bulletin*, 81(12):899, 1974. [Page 14]

[68] R. Metzler and J. Klafter. The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Physics Reports*, 339(1):1–77, 2000. [Page 83]

[69] Y. Nagai. From bottom-up visual attention to robot action learning. In *Proceedings of 8 IEEE International Conference on Development and Learning*. IEEE Press, 2009. [Page 84]

[70] J. Najemnik and W.S. Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, 2005. [Pages 15 and 78]

[71] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision research*, 45(2):205–231, 2005. [Page 77]

[72] Vidhya Navalpakkam, Christof Koch, Antonio Rangel, and Pietro Perona. Optimal reward harvesting in complex perceptual environments. *Proceedings of the National Academy of Sciences*, 107(11):5232–5237, 2010. [Pages 25 and 36]

[73] Vladimir Nedovic, Arnold WM Smeulders, Andre Redert, and J-M Geusebroek. Stages as models of scene geometry. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1673–1687, 2010. [Page 19]

[74] J.P. Nolan. Numerical calculation of stable densities and distribution functions. *Communications in Statistics-Stochastic Models*, 13(4):759–774, 1997. [Page 83]

[75] A. Oliva, A. Torralba, M.S. Castelhano, and J.M. Henderson. Top-down control of visual attention in object detection. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 1, pages I–253. IEEE, 2003. [Page 21]

[76] J.K. O'Regan. Solving the" real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 46(3):461, 1992. [Page 17]

[77] J.K. O'Regan, A. Noë, et al. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5):939–972, 2001. [Page 17]

[78] J. Otero-Millan, X.G. Troncoso, S.L. Macknik, I. Serrano-Pedraza, and S. Martinez-Conde. Saccades and microsaccades during visual fixation, exploration, and search: foundations for a common saccadic generator. *Journal of Vision*, 8(14), 2008. [Page 78]

[79] Luiz Pessoa. On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, 9(2):148–158, 2008. [Page 78]

[80] Luiz Pessoa and Ralph Adolphs. Emotion processing and the amygdala: from a'low road'to'many roads' of evaluating biological significance. *Nature Reviews Neuroscience*, 11(11):773–783, 2010. [Page **78**]

[81] Matthew S Peterson, Arthur F Kramer, Ranxiao Frances Wang, David E Irwin, and Jason S McCarley. Visual search has memory. *Psychological Science*, 12(4):287–292, 2001. [Page **31**]

[82] MJ Plank and A. James. Optimal foraging: Lévy pattern or process? *Journal of The Royal Society Interface*, 5(26):1077, 2008. [Page **85**]

[83] H.K. Preisler, A.A. Ager, B.K. Johnson, and J.G. Kie. Modeling animal movements using stochastic differential equations. *Environmetrics*, 15(7):643–657, 2004. [Page **84**]

[84] Graham H Pyke, H Ronald Pulliam, and Eric Charnov. Optimal foraging: a selective review of theory and tests. *Quarterly Review of Biology*, 52:137–154, 1977. [Page **81**]

[85] Rajesh P.N. Rao, Gregory J. Zelinsky, Mary M. Hayhoe, and Dana H. Ballard. Eye movements in iconic visual search. *Vision Research*, 42(11):1447 – 1463, 2002. [Page **77**]

[86] R.P.N. Rao. Bayesian inference and attentional modulation in the visual cortex. *Neuroreport*, 16(16):1843, 2005. [Page **22**]

[87] R. Ratcliff and G. McKoon. The diffusion decision model: Theory and data for two-choice decision tasks. *Neural computation*, 20(4):873–922, 2008. [Page **84**]

[88] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998. [Pages **7**, **14** and **15**]

[89] K. Rayner. Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, 62(8):1457–1506, 2009. [Page **15**]

[90] R.A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 1(3):17–42, 2000. [Pages **7**, **17**, **34** and **39**]

[91] R.A. Rensink, J.K. O'Regan, and J.J. Clark. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8(5):368, 1997. [Pages **3**, **7**, **17** and **18**]

[92] A. Reynolds. How many animals really do the Lévy walk? Comment. *Ecology*, 89(8):2347–2351, 2008. [Page **85**]

[93] AM Reynolds. Optimal random Lévy-loop searching: New insights into the searching behaviours of central-place foragers. *EPL (Europhysics Letters)*, 82:20001, 2008. [Page **85**]

[94] I. Rhee, M. Shin, S. Hong, K. Lee, S.J. Kim, and S. Chong. On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking,*, 19(3):630–643, 2011. [Page **48**]

[95] H. Risken. *The Fokker-Planck equation: Methods of solution and applications*. Springer–Verlag, Berlin, Germany, 1996. [Page **83**]

[96] C.A. Rothkopf, D.H. Ballard, and M.M. Hayhoe. Task and context determine where you look. *Journal of Vision*, 7(14), 2007. [Pages **17**, **22**, **25** and **36**]

[97] A.C. Schütz, D.I. Braun, and K.R. Gegenfurtner. Eye movements and perception: A selective review. *Journal of Vision*, 11(5), 2011. [Pages **24** and **78**]

[98] A. Shahab, F. Shafait, A. Dengel, and S. Uchida. How salient is scene text? In *Proc. 10th IAPR International Workshop onDocument Analysis Systems (DAS, 2012)*, pages 317–321. IEEE, 2012. [Page **39**]

[99] Satoshi Shioiri and Mitsuo Ikeda. Useful resolution for picture perception as a function of eccentricity. *Perception*, 18:347–361, 1989. [Page **54**]

[100] M.F. Shlesinger and J. Klafter. *Lévy walks versus Lévy flights*. Martinus Nijhof Publishers, Amsterdam, 1986. [Page **84**]

[101] Daniel J Simons. Current approaches to change blindness. *Visual cognition*, 7(1-3):1–15, 2000. [Pages **7** and **17**]

[102] GW Snedecor and WG Cochran. *Statistical methods*. Iowa State University Press, Ames, IA, 8-th edition, 1989. [Page **57**]

[103] Alec Solway and Matthew M Botvinick. Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychological review*, 119(1):120, 2012. [Page **35**]

[104] Nathan Sprague and Dana Ballard. Eye movements for reward maximization. In *Advances in neural information processing systems*, volume 16. MIT Press, Cambridge, MA, 2003. [Page **80**]

[105] Andreas Sprenger, Monique Friedrich, Matthias Nagel, Christiane S Schmidt, Steffen Moritz, and Rebekka Lencer. Advanced analysis of free visual exploration patterns in schizophrenia. *Frontiers in psychology*, 4, 2013. [Page **78**]

[106] D.G. Stephen, D. Mirman, J.S. Magnuson, and J.A. Dixon. Lévy-like diffusion in eye movements during spoken-language comprehension. *Physical Review E*, 79(5):056114, 2009. [Page **84**]

[107] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5), 2011. [Page **54**]

[108] Yaoru Sun, Robert Fisher, Fang Wang, and Herman Martins Gomes. A computer vision model for visual-object-based attention and eye movements. *Computer Vision and Image Understanding*, 112(2):126 – 142, 2008. [Pages **44** and **77**]

[109] B.W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 2007. [Page **23**]

[110] B.W. Tatler, M.M. Hayhoe, M.F. Land, and D.H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5), 2011. [Pages **23**, **26**, **33**, **47**, **48**, **50** and **58**]

[111] B.W. Tatler and B.T. Vincent. The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17(6-7):1029–1054, 2009. [Pages **23** and **26**]

[112] B.W. Tatler and B.T. Vincent. The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17(6-7):1029–1054, 2009. [Pages **47**, **48** and **78**]

[113] S. Thorpe, D. Fize, C. Marlot, et al. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996. [Page **25**]

[114] A Torralba. Contextual priming for object detection. *Int. J. of Comp. Vis.*, 53:153–167, 2003. [Pages **21**, **32**, **34**, **39** and **79**]

[115] A. Torralba. Modeling global scene factors in attention. *JOSA A*, 20(7):1407–1418, 2003. [Page **21**]

[116] A. Torralba, A. Oliva, M.S. Castelhano, and J.M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006. [Page **21**]

[117] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. [Pages **16** and **20**]

[118] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. [Page **7**]

[119] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009. [Page **78**]

[120] Paul Viola and MichaelJ. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. [Page **80**]

[121] GM Viswanathan, EP Raposo, and MGE da Luz. Lévy flights and superdiffusion in the context of biological encounters and random searches. *Physics of Life Rev.*, 5(3):133–150, 2008. [Page **84**]

[122] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006. [Pages **25**, **32**, **34** and **42**]

[123] H.C. Wang and M. Pomplun. The attraction of visual attention to texts in real-world scenes. *Journal of Vision*, 12(6), 2012. [Page **22**]

[124] H.C. Wang and M. Pomplun. The attraction of visual attention to texts in real-world scenes. *Journal of Vision*, 12(6), 2012. [Pages **32**, **36** and **52**]

[125] A. Weron and R. Weron. Computer simulation of lévy $\alpha$-stable variables and processes. In P. Garbaczewski, M. Wolf, and A. Weron, editors, *Chaos –The Interplay Between Stochastic and Deterministic Behaviour*, volume 457 of *Lecture Notes in Physics*, pages 379–392. Springer Berlin / Heidelberg, 1995. [Page **83**]

[126] Niklas Wilming, Simon Harst, Nico Schmidt, and Peter König. Saccadic momentum and facilitation of return saccades contribute to an optimal foraging strategy. *PLoS Comput. Biol.*, 9(1):e1002871, 2013. [Pages **34** and **78**]

[127] Marco Wischnewski, Anna Belardinelli, Werner Schneider, and Jochen Steil. Where to Look Next? Combining Static and Dynamic Proto-objects in a TVA-based Model of Visual Attention. *Cognitive Computation*, 2(4):326–343, 2010. [Pages **32** and **34**]

[128] Jeremy M Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994. [Page **77**]

[129] Jeremy M. Wolfe. When is it time to move to the next raspberry bush? foraging rules in human visual search. *Journal of Vision*, 13(3), 2013. [Page **8**]

[130] J.M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994. [Page **21**]

[131] J.M. Wolfe. Guided search 4.0. *Integrated models of cognitive systems*, pages 99–119, 2007. [Page **21**]

[132] A.L. Yarbus. *Eye Movements and Vision*. Plenum Press, New York, 1967. [Pages **16** and **22**]

[133] Gregory J Zelinsky. A theory of eye movements during target acquisition. *Psychological review*, 115(4):787, 2008. [Pages **36** and **77**]

[134] L. Zhang, M.H. Tong, and G.W. Cottrell. Sunday: Saliency using natural statistics for dynamic analysis of scenes. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 2944–2949, 2009. [Page **21**]

[135] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008. [Page **21**]

[136] Xin Zhang, Yee-Hong Yang, Zhiguang Han, Hui Wang, and Chao Gao. Object class detection: A survey. *ACM Comput. Surv.*, 46(1):10:1–10:53, July 2013. [Page **7**]

[137] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of vision*, 11(3), 2011. [Page **84**]

# Publications

## Journal Articles:

- **Antonio Clavelli**, Dimosthenis Karatzas, Josep Lladòs, Mario Ferraro, Giuseppe Boccignone, "Modelling Task-Dependent Eye Guidance to Objects in Pictures", *Cognitive Computation, (Springer)*, 2014

## Conferences

- **Antonio Clavelli**, Dimosthenis Karatzas, Josep Lladòs, Mario Ferraro, Giuseppe Boccignone, "Towards modelling an attention-based text localization process" In *6th Iberian conference on Pattern recognition and image analysis (IbPRIA 2013)*, Funchal, Madeira, Portugal, 2013, pp 296–303.
- **Antonio Clavelli**, Dimosthenis Karatzas, Josep Lladòs, "A framework for the assessment of text extraction algorithms on complex colour images" In *9th International Workshop on Document Analysis Systems (DAS 2010)*, Boston, MA, 2010, pp 19–28.
- **Antonio Clavelli** and Dimosthenis Karatzas, "Text Segmentation in Colour Posters from the Spanish Civil War Era" In *10th International Conference Document Analysis and Recognition (ICDAR 2009)*, Barcelona, Spain, 2009, pp 181–185.