UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

DIPARTIMENTO DI INFORMATICA 'GIOVANNI DEGLI ANTONI'
CORSO DI DOTTORATO IN INFORMATICA
XXXIII CICLO

TESI DI DOTTORATO DI RICERCA

A STOCHASTIC FORAGING MODEL OF ATTENTIVE EYE GUIDANCE ON DYNAMIC STIMULI

SSD 01/B1

Autore
Alessandro D'Amelio

Supervisor
Dott. Giuliano Grossi

Co-Supervisor
Prof. Giuseppe Boccignone

Coordinatore del Dottorato
Prof. Paolo Boldi

A.A. 2019-2020

...this work is dedicated to my family.

Acknowledgments

"Data Science is an eyeglass enabling us to look through the window of reality. It is not a mirror through which data looks at itself under make-up"

Judea Pearl

First and foremost, I would like to thank my Supervisor Dr. Giuliano Grossi, for his great willingness and for having inspired me since my early Master degree studies; it is by and large his merit (or fault) if I decided to pursue a career as a PhD student. I am particularly grateful to my Co-Supervisor Prof. Giuseppe Boccignone, his advises and wisdom have been crucial for this thesis, from its definition up to the writing of the dissertation. More importantly, he has taught me how powerful and interesting may be to look at problems from different perspectives. I would also like to express my very great appreciation to Dr. Raffaella Lanzarotti for her wide availability and great help in taking my first steps as a researcher.

My special thanks are extended to all my mates at the PHuSe Lab since I joined it: Vittorio Cuculo, Sathya Bursic, Marco Granato, Claudio Ceruti and Sabrina Patania, with special reference to Vittorio, the *senior* member; I've always had the opportunity to share with him any doubt I've ever had. Everything is much simpler when you have some footprints to follow. I also want to thank Sathya for all the discussions we had on Machine Learning concepts, for the guitar/music tips we shared and for having taught me few words in Croatian.

I wish to acknowledge Dr. Tom Foulsham for having hosted me at the University of Essex for six beautiful weeks and for the enlightening discussions on eye movements.

A big thank to all my friends that despite the distance and the few opportunities to meet, remain an anchor in my life.

I'm eternally grateful to my family, in particular Antonio and Pierangela for their constant support and having allowed me to find my own way, and to my sister Sabrina, it was great to spend the last year as housemates, like we were kids. Many thanks to Giovanni and Rosetta for always having manifested their pride about what I was doing.

Last but certainly not least, thanks to my life companion Alessandra, for always being by my side and so sympathetic.

Abstract

UNDERSTANDING human behavioural signals is one of the key ingredients of an effective human-human and human-computer interaction (HCI). In such respect, non verbal communication plays a key role and is composed by a variety of modalities acting jointly to convey a common message. In particular, cues like gesture, facial expression, prosody etc. have the same importance as spoken words. Gaze behaviour makes no exception, being one of the most common, yet unobtrusive ways of communicating.

To this aim, many computational models of visual attention allocation have been proposed; although such models were primarily conceived in the psychological field, in the last couple of decades, the problem of predicting attention allocation on a visual stimuli has started to catch the interest of the computer vision and pattern recognition community, pushed by the fast growing number of possible applications (e.g. autonomous driving, image/video compression, robotics).

In this renaissance of attention modelling, some of the key features characterizing eye movements were at best overlooked; in particular the explicit unrolling in time of eye movements (i.e. their dynamics) has been seldom taken into account. Moreover, the vast majority of the proposed models are only able to deal with static stimuli (images), with few notable exceptions.

The main contribution of this work is a novel computational model of attentive eye guidance which derives gaze dynamics in a principled way, by reformulating attention deployment as a stochastic foraging problem. We show how treating a virtual observer attending to a video as a stochastic composite forager searching for valuable patches in a multi-modal landscape, leads to simulated gaze trajectories that are not statistically distinguishable from the ones performed by humans while free-viewing the same scene.

Model simulation and experiments are carried out on a publicly available dataset of eye-tracked subjects displaying conversations and social interactions between humans.

Estratto

La comprensione dei segnali comportamentali umani è uno dei fattori principali di una efficiente interazione uomo-uomo e uomo-macchina. Da questo punto di vista, la comunicazione non verbale gioca un ruolo fondamentale ed è composta da una varietà di modalità che agiscono congiuntamente con l'obiettivo di veicolare un messaggio comune. Nello specifico, segnali come la gestualità, l'espressione facciale, la prosodia ecc, hanno la stessa importanza del parlato. Il comportamento dello sguardo non fa eccezione, essendo uno dei meccanismi più comuni sebbene uno dei più discreti per comunicare.

Questa è la finalità principale che ha portato allo sviluppo di diversi modelli computazionali di allocazione dell'attenzione visiva; sebbene questi siano stati originariamente di interesse strettamente relativo all'area psicologica, negli ultimi vent'anni il problema relativo alla predizione dell'attenzione su uno stimolo visivo, ha iniziato ad attrarre l'interesse delle comunità della visione artificiale e del riconoscimento dei pattern, spinte dal crescente numero di possibili applicazioni (dalla guida autonoma alla compressione di immagini e video, fino alla robotica).

In questo rinascimento della modellazione dell'attenzione, alcune caratteristiche fondamentali tipiche dei movimenti oculari sono state, nel migliore dei casi, trascurate; in particolare, la definizione della dinamica temporale dei movimenti oculari è stata raramente presa in considerazione. Inoltre, la stragrande maggioranza dei modelli proposti in letteratura, si limita all'analisi di stimoli statici (immagini), con rare eccezioni.

Il contributo principale di questo lavoro è un nuovo modello computazionale di attenzione visiva che deriva la dinamica dello sguardo da fermi principi. Il problema dell'allocazione attentiva viene quindi riformulato come un problema di foraging stocastico. Viene altresì mostrato come, trattare un generico osservatore come un *forager composito stocastico* alla ricerca di zone ricche all'interno di un ambiente multimodale, permetta di simulare delle traiettorie dello sguardo che non sono distinguibili a livello statistico da quelle compiute da umani che osservano la medesima scena. Simulazioni del modello proposto e relativi esperimenti, sono eseguiti su un dataset pubblico contenente il tracciamento oculare di soggetti che osservano interazioni sociali tra persone.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

As humans, we are immersed in a multitude of sensory data delivered to our senses as auditory, visual, tactile signals etc. Such amount of information is too large to be entirely processed at any time; as a consequence most valuable information must be spotted and prioritized. Evolution has solved this problem by equipping us with the mechanism of *selective attention*; we are able to circumscribe the kind of information we are interested in, in order to reduce the perceived complexity of the surrounding environment. This ability exists for each of our senses. For example, in the field of auditory attention it's worth mentioning the well known *cocktail party effect*: in a room full of different voices and sounds, we are able to focus on a particular voice of a certain person (Cherry, 1953).

Likewise, visual attention has its own way of being "selective". The *fovea*, the center of the retina, is the region with the highest resolution of the eye. Our oculomotor system allows us to continuously move our eyes in order to keep inside the fovea the region of the visual landscape with the highest interest to the observer, thus automatically attributing less "descriptive power" to the remaining part of the stimuli. This fact has been revealed by various experiments on *change blindness* (Simons and Levin, 1997) in which significant changes in the scene are not noticed by observers (observers are "blind" for such changes).

In order to provide the impression of retaining a rich representation of the surrounding world, phases of visual intake (fixations) are followed large relocations (saccades) towards other regions of the stimuli. Such pattern has been referred to as a "saccade and fixate" strategy (Land, 2006). Saccades are the fast movements that redirect the eye to a new part of the surroundings, and fixational movements occur within intervals between saccades, in which gaze is held almost stationary. In dynamic scenes, or ones including observer's movement, fixations are either replaced by or augmented with the smooth

pursuit eye movement to keep on the fovea the objects of interest that are moving.

However, the sequence of eye movements directing the focus of attention to a specific region of interest only tell a part of the story; it has been known since Von Helmholtz (1867) about the ability of attending a particular location of the visual field without performing any eye movement:

> *"I found myself able to choose in advance which part of the dark field off to the side of the constantly fixated pinhole I wanted to perceive by indirect vision"*

Such ability is called *covert attention* in contrast to the explicit selection of specific regions of the visual landscape through eye movements called *overt attention*. Crucially, there has been given evidence (Deubel et al., 1996) that both such mechanisms work together in order to perform complex vision tasks.

Interestingly enough, the information coming from different senses may be combined in order to have a more effective selection mechanism. Mutual influence between speech and visual perception, markedly, face perception, is a long debated and well known issue. The link between perceiving speech and perceiving faces has been demonstrated in both behavioural and physiological experiments, e.g., (McGurk and MacDonald, 1976; Sumby and Pollack, 1954; Ross et al., 2007; Calvert et al., 1997; Kriegstein et al., 2005).

The McGurk effect (McGurk and MacDonald, 1976) is one celebrated example of audio-visual speech perception, where visual inputs can even override the veridical inputs of the auditory system. Another example is the way people routinely use information provided by the speaker's lip movements to help understand speech in a noisy environment (Sumby and Pollack, 1954; Ross et al., 2007). Watching the lips move in silent video clips activates areas in the auditory cortex that are activated when people are perceiving speech (Calvert et al., 1997); conversely, when listeners pay attention to a voice that they associate with a specific person (Kriegstein et al., 2005), this activates areas not only for perceiving speech but also for perceiving faces (face fusiform area, FFA). Van der Burg et al. (2008) provided evidence that that audio-visual synchrony guides attention in an exogenous manner in adults.

It has been argued that similarities between auditory and visual perception in complex scenes suggest that common neural mechanisms control attention across modalities (Shinn-Cunningham, 2008). However, it remains unclear how multimodal scenes are represented in the brain (Kondo et al., 2017) and there is no comprehensive framework to explain our abilities in multimodal attention.

At this point a series of questions may arise: *"what's the kind of mechanism that drives our attention towards a particular region of the audio-visual landscape?"* or *"How much time do we need to spend looking at that region before directing the gaze to the next one?"*, *"How do I decide where to look next?"* and *"Can this process be eventually rigorously described?"* In essence, all these queries denote a quest for describing the dynamics of the audio-visual attentive process. This is the chief intent of the present work, in which such account is made by means of the development of a computational model of visual attention.

We rely our investigation on videos displaying conversations and social interactions between people. Indeed, conversational videos have the ecological virtue of displaying

real people embedded in a dynamic situation while being relatively controlled stimuli. Besides that, this kind of data is, nowadays, ubiquitous due to the recent rise of dedicated channels (Truong and Agrawala, 2019; Pires and Simon, 2015). Therefore, it will be used (together with recordings of eye movements of observers attending at such stimuli) as a test-bed to validate the modelling assumptions.

Throughout the dissertation, a stochastic foraging perspective on eye movements will be presented and adopted not only as an informing metaphor, but rather as a sound framework for modelling gaze deployment to audio-visual dynamic stimuli. Indeed, animals searching for food in a patchy landscape and eyes wandering on a visual stimuli tend to yield similar patterns (Brockmann and Geisel, 2000). Interestingly enough, Ecological literature has developed a theoretical basis to answer most of the questions listed above, in the context of foraging animals. For instance, it has been shown that the dynamics of animal movement can be modelled by stochastic processes of the Lèvy type (Viswanathan et al., 1996). However, later developments have argued that the same behaviour is well described by a mixture of classical brownian walks (Benhamou, 2007). In this work, we take advantage of such body of knowledge in order to formulate an explainable model of human's visual attention mechanisms. As a natural outgrowth, stochastic processes will be employed in order to provide a rigorous description of eye movements. Results, obtained on a publicly available dataset, prove the efficacy of such simile.

The thesis is organized as follows:

Chapter 2 gives a broad overview of the early approaches to computational modelling of attentive eye guidance, together with most recent developments, mainly broad by the computer vision community. In the same chapter, a criticism to some modern modelling practices is made and a novel model for the prediction of sequences of eye movements on static images is presented.

Chapter 3 lays the basis of stochastic processes for the description of eye movements, introduces the foraging perspective and how it can be applied to model human's oculomotor behaviour.

Chapter 4 presents a novel computational model of gaze deployment to audio visual stimuli of social interactions, inspired by ecological models of foraging animals. In particular the human observer is modeled as a stochastic composite forager searching for valuable informative patches in a dynamic visual landscape. The patch choice and residence times are derived from principles of Optimal Foraging Theory. The dynamics of eye movements is described at the tiniest scale via a Stochastic Differential Equation (SDE) with switching parameters.

Chapter 5 provides simulations, and an in depth statistical assessment of results. In particular an ablation study of the proposed model is conducted in order to validate the modelling assumptions, together with a comparison with state of the art gaze control models. Remarkably, evidence is given of the non-discernibility between model simulated gaze data and real eye movements as recorded from eye trackers, in a statistical sense.

Chapter 6 summarizes the key contributions of the thesis and presents some concluding remarks.

CHAPTER 2

Computational Models of Attentive Eye Guidance

ACCORDING to the braod definition of attention given by Corbetta (1998): *"Attention defines the mental ability to select stimuli, responses, memories, or thoughts that are behaviorally relevant among the many others that are behaviorally irrelevant"*. Crucially, such definition in the realm of visual attention, poses the necessity of bringing into the game the concept of *relevance* of cues under specific tasks or goals of the observer (behavioral relevance) . In other words, the attention allocation does not uniquely depends on the stimuli, but it's rather influenced by the internal state of the subject.

As stated previously, this relentless information picking is accomplished through a sequence of eye movements (overt attention). Such behaviour is conditioned by a covert attentive selection mechanism based on task knowledge and goals of the observer in order to enhance the perception process. It has been given evidence that covert attentive mechanisms play an important role in guiding overt orienting based on eye movements (Hoffman, 1998). It appears that explicit eye movements towards a specific region are preceded by shifts of attention to the same location.

Moreover this coupling persists regardless of whether the eye movement is triggered by bottom-up factors (sudden movements) or top-down influences like endogenous control, instructions or expectations (Hoffman, 1998). In other words, if the overt attention mechanisms entail the explicit action of selecting a particular region of interest, the definition of which regions of the stimuli can be marked as interesting is entrusted to the covert attention mechanisms that are mediated by the goals (either internal or external) of the observer.

Early approaches to modelling the gaze shift behaviour have their roots in the pioneering work on active vision (Aloimonos et al., 1988; Ballard, 1991) which integrate the problem of of vision into an action-perception loop in which the sensory apparatus

of the organism acts actively on the environment, for instance by manipulating the view point of the camera (action), in order to efficiently sample it (perception).

Such approaches as remarked by Rothenstein and Tsotsos (2008) overlooked the link between gaze shifts (overt attention) and covert attention defined as the ability of selection based on task knowledge and observer's goals. The latter aspects, according to Rothenstein and Tsotsos (2008) accounts the observer to a powerful heuristic to limit search and make the overall problem of "search for information" tractable in terms of computational complexity.

Under such rationale, a sound computational model of visual attention should take into account both the eye movements and all such mechanisms of attentive selection based on task knowledge that are practically carried out in *pre-attentive* computations. These involve saliency, plans, objects and values (Schütz et al., 2011).

Such mechanisms require a distinction between so called *bottom-up* (saliency) and *top-down* (plans, objects, values) attentional cues. Bottom-up factors are derived solely from the visual scene in the form of features that induce a "reaction" of the observer (sudden movements or sounds, high contrast regions etc...); such areas of the stimuli are called *salient* and are usually represented spatially through *saliency maps*.

On the other hand, top-down attention is driven by cognitive factors such as knowledge, expectations and current goals; Early evidence for top-down attentional control where given by Yarbus (1967) which showed that, for the same scene, observers with different tasks ("estimate age of people" vs. "estimate the material circumstances of people" or simply "look freely") would produce very different eye movements.

All such attentional control mechanisms work jointly in order to define which parts of the stimuli are important, thus telling where to look and in which order. Computational modelling of visual attention has to do with the explicit description of of this complex machinery: covert and overt attention, bottom-up vs. top-down control, spatial salience and the unfolding in time of eye movements.

The following sections aim at providing a general overview of the computational models of attentive eye guidance that have been presented in the scientific literature. Starting from the earliest approaches, mainly rooted in the computational psychology field we will move to more recent ones tackled as computer vision and pattern recognition issues. Some of these models will be briefly described, discussing the main hurdles concerning the modelling of gaze deployment together with some criticism; in this vein, new experimental results will be provided.

2.1 Overview of early approaches

The problem of modeling visual attention has been initially tackled from the psychological point of view. Indeed, the psychological literature presents a variety of theories and models in this compound aiming at understanding human perception (Frintrop et al., 2010).

In such vein, one of the most influential and known model is the Treisman's *Feature Integration Theory* (*FIT*) (Treisman and Gelade, 1980; Treisman, 1998); according to FIT, different features are selected across the visual field in parallel at an early stage, while objects are identified later and separately. Such features, are represented by different *feature maps* that are later fused to yield a *master map* which is the topographical

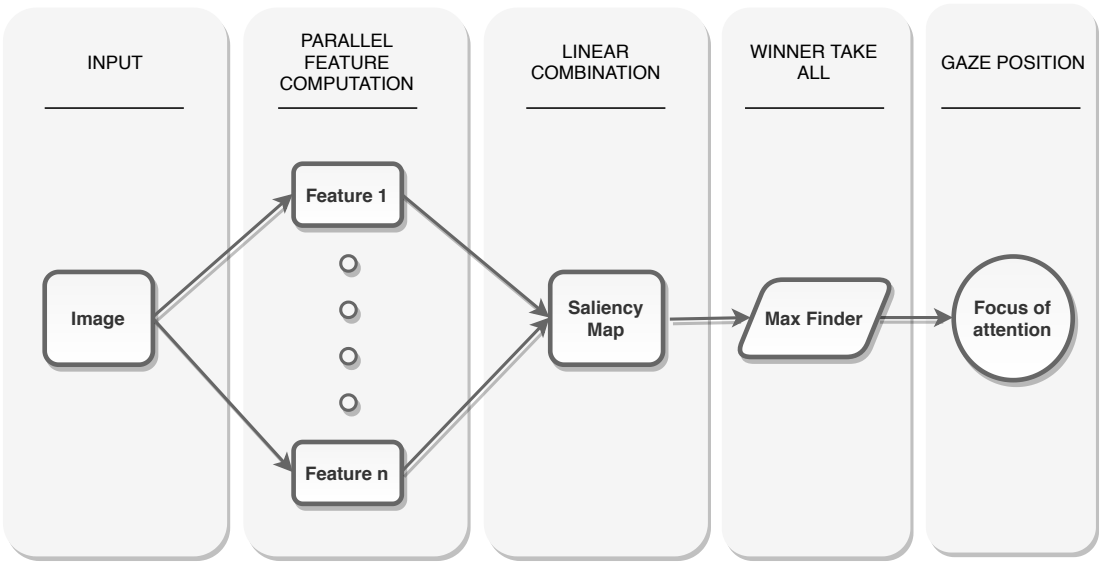


Figure 2.1: *General structure of a computational model of visual attention*

representation of all the features.

Similarly to the Treisman's model, Wolfe (1994) proposed the *Guided Search Model*, another popular model which shares with the FIT many of the concepts and the architectural design.

Beside these approaches, there is a wide variety of psychophysical models on visual attention. Without wanting to be exhaustive, some notable example are the *Biased Competition Model* by Desimone and Duncan (1995), the *Koch and Ullman's model* (Koch and Ullman, 1985) and Tsotsos' *Selective Tuning* model (Tsotsos et al., 1995). At a different level of explanation, other proposals have been conceived in terms of connectionist models, such as *MORSEL* (Multiple Object Recognition and attentional SELection, (Mozer, 1987)), *SLAM* (SeLective Attention Model, (Phaf et al., 1990)), *SERR* (SEarch via Recursive Rejection, (Humphreys and Muller, 1993)), and *SAIM* (Selective Attention for Identification Model by Heinke and Humphreys (2003)) subsequently refined in the Visual Search *SAIM* (VS-SAIM) (Heinke and Backhaus, 2011).

All such theoretical models, conceived in the psychological field, typically deal with very simple stimuli like synthetical images. The goal of computational vision is to deliver models that are eventually able to deal with more complex stimuli like natural scenes. To this end, in the last 15-20 years a number of models of visual attention have been proposed.

The general structure of all such models can be described as in Figure 2.1; despite each system may vary in detail, most of them share such similar structure. A notable example is the model by Itti et al. (1998), which is probably the most popular computational model of visual attention. Since its publication, the field flourished.

The model is the computational counterpart of the Koch and Ullman (1985) and Treisman's FIT models and relies on the computation of several features in parallel that are then collected in maps that may be represented as gray scale images whose brightness is proportional to the intensity of the feature at hand (color, intensity, orientation) and are often called *conspicuity maps*. Such features are computed by a set of

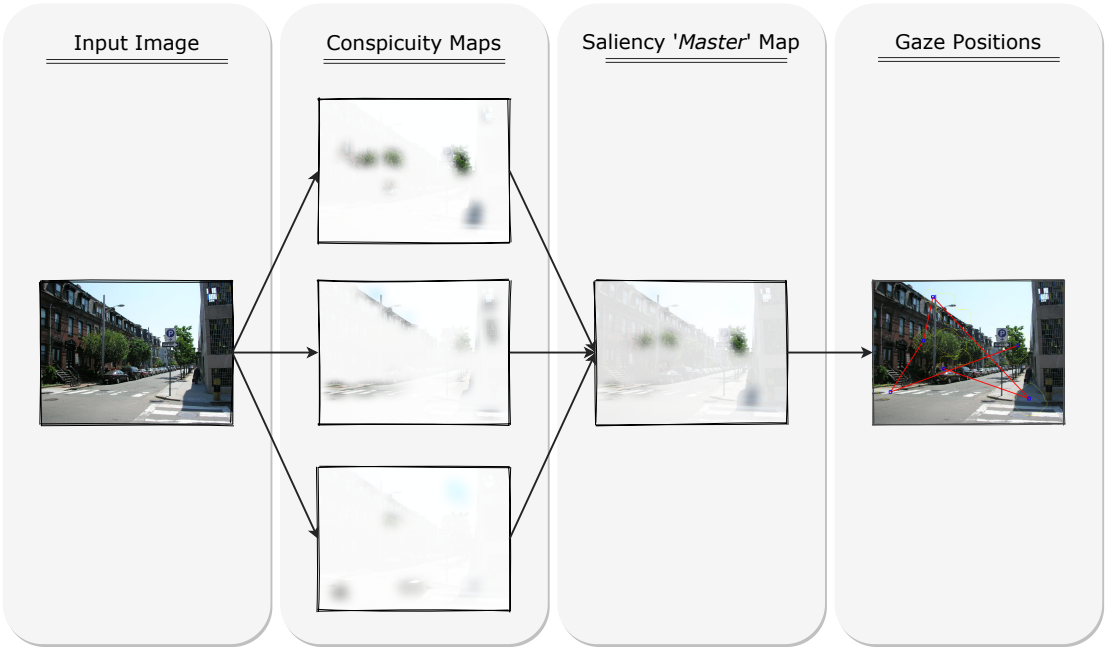


Figure 2.2: *Sample output of the model proposed by Itti et al. (1998)*

linear *center surround* operations, mimicking the mechanism of visual receptive fields. To this end, prior to feature computation, the input image is re-sampled at 8 different scales via gaussian pyramids; center-surround is thus implemented as the difference between fine and coarse scales. This allows detect locations which stand out from their surround as happens in the retina, lateral geniculate nucleus, and primary visual cortex.

The conspicuity maps are then fused to a single saliency map (or Treisman's master map of location) via linear combination. The sequence of fixation location on the input image is then determined by selecting the local maxima of the saliency map through a *winner-take-all* (WTA) network, with transient inhibition of fixated locations to avoid the model becoming stuck on the same portion of the stimuli, a mechanism usually termed "inhibition of return" (IOR). Figure 2.2 depicts the output of such procedure on a sample image.

This approach is strongly biologically motivated and shows how such a mechanism might work in the human brain (Koch and Ullman, 1985).

Crucially, the sequential WTA selection mechanism allow the model produce the sequence of fixation and saccades given a particular stimuli that mimic the overt attentional control of humans' visual system. Indeed, our oculomotor control mechanism is in charge of deciding, at any time, which portion of the scene is worth choosing; hence, in a crude summary, the aim of a computational model mimicking attentive eye guidance boils down to answering the question: *Where to look next?*

By further abstracting the structure of the model given in Figure 2.1 one such question can be practically addressed by providing an account of the mapping from visual data of a natural scene, say $\mathbf{I}$ (raw image data representing either a static picture or a stream of images), to a sequence of time-stamped gaze locations $(\mathbf{r}_{F_1}, t_1), (\mathbf{r}_{F_2}, t_2), \ldots$, namely (Boccignone, 2016):

$$\mathbf{I} \rightarrow \{\mathbf{r}_{F_1}, t_1; \mathbf{r}_{F_2}, t_2; \dots\} \tag{2.1}$$

From the modelling standpoint, given a stimuli $\mathbf{I}$, either static (image) or dynamic (video) the only observations that are given are the sequence of locations of the scene visited by the observer. In case of static stimuli the sequece of continuous eye movements can be classified into the corresponding sequence of fixations and saccades; conversely, when dealing with videos, smooth pursuits should be taken into account. Here we adopt the generic term of *gaze shift* to describe a sequence of either pursuits, saccades or fixations. For the sake of notational simplicity, from now on, we will write the time series $\{\mathbf{r}_{F_1}, t_1; \mathbf{r}_{F_2}, t_2; \dots\}$ as $\{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}$, thus adopting the compact notation: $(\mathbf{r}_{F_n}, t_n) = \mathbf{r}_F(n)$.

The common practice to derive the mapping 2.1, is to conceive it as a two step approach:

(a) Compute the perceptual representation:

$$\mathbf{I} \rightarrow \mathcal{W} \tag{2.2}$$

(b) use $\mathcal{W}$ to generate the scanpath:

$$\mathcal{W} \rightarrow \{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\} \tag{2.3}$$

Note how Figure 2.2 can be effectively described by this procedure.

By and large, recent literature in the field of computational modeling has been mainly concerned with the first step, i.e. deriving a perceptual representation, typically in the form of a saliency map. This allowed to produce models able to answer the question of *where* to look at on a given stimuli, thus putting aside the temporal dimension of the gaze deployment process. In the meanwhile, the second step answering the question on *How* to look at it has been often overlooked. A deeper treatment of both such questions is provided in the following sections.

2.2 Perceptual Representation: The Salience Conundrum

As stated earlier, building the perceptual representation of the stimuli, involves the selection of *Where* to gaze at - features, objects, actions - and their location within the scene. By and large (Tatler et al., 2011; Borji and Itti, 2013; Bruce et al., 2015; Bylinskii et al., 2015), the computational modelling of visual attention has hitherto been concerned with this particular aspect of the wider framework described earlier: deriving a representation $\mathcal{W}$. This is, in essence, the aim of the multitude of so called *saliency models*.

2.2.1 Saliency Models

In a nutshell, *saliency models* are algorithms that take an image $\mathbf{I}(\mathbf{r})$ as input, and return topographic maps $\mathcal{W}(\mathbf{r})$ indicating the salience at each location $\mathbf{r} = (x, y)$ in the image (the likelihood of fixating at $\mathbf{r}$). This models are appealing since, apparently, they represent a straightforward operational definition of visual attention - the allocation of

visual resources to the viewed scene (Bylinskii et al., 2015). As a consequence they have gained currency for a variety of applications in computer vision, image and video processing and compression, quality assessment (Nguyen et al., 2018). Indeed they put the accent on the importance of particular cues inside the image/video, thus allowing such models to deviate from the original role of tool employed to understand human attentive behaviour (as conceived in the computational psychology field) but becoming a way to process the image itself. This fact is particularly evident in modern deep saliency models in which the black box nature of such architectures prevents the model from any explanatory purpose.

In the celebrated model proposed by Itti et al. (1998), the mapping $\mathbf{I} \rightarrow \mathcal{W}$ was performed via computation of a set of *conspicuity maps*, later fused into a *saliency map*, by relying on simple features of the image (color, intensity and orientation computed via gabor pyramids). Such approach can be easily recognized as a *bottom-up* one: it is assumed that the attention of the observer is mainly captured by the low level cues of the scene. At least in the early implementation, this model does not take into account the *top-down* information.

There has been a long debated controversy concerning the bottom-up vs. top-down nature of eye guidance control (Egeth and Yantis, 1997; Tatler et al., 2011), however recent studies and empirical evidence, suggest that factors such as context (Torralba et al., 2006), spatial biases (Tatler and Vincent, 2009), affect and personality (Cuculo et al., 2018), dynamics of attention deployment (Tatler et al., 2005; Schütt et al., 2019) are likely to play a key role and might contribute in subtle ways to effectiveness and performance of saliency models (Tatler et al., 2011; Kummerer et al., 2017; Kong et al., 2018; Schütt et al., 2019).

Nonetheless, up to this date, as stigmatised in many studies (Foulsham and Underwood, 2008; Einhäuser et al., 2008; Tatler et al., 2011; Borji and Itti, 2013; Bruce et al., 2015; Bylinskii et al., 2015), the majority of computational models have retained a central place for low-level visual conspicuity without referring to the semantic content of the scene.

The weakness of the bottom-up approach has been largely weighed up in the visual attention realm (Tatler et al., 2011; Foulsham and Underwood, 2008; Einhäuser et al., 2008; Schütz et al., 2011); indeed, it has been argued that early salience has only an indirect effect on attention by acting through recognized objects: observers attend to interesting objects and salience contributes little extra information to fixation prediction (Einhäuser et al., 2008). Moreover Schütz et al. (2011) argued on the plausibility of a multitude on representational levels to account for: 1) *salience*, 2) *objects*, 3) *values*, and 4) *plans*.

To overcome this pitfall, early saliency can be top-down tuned to improve its fixation prediction when dealing with objects, faces, text regions or contextual cues. In this approaches, besides the bottom-up information (saliency), the set of objects of interest at hand is determined; for example, Cerf et al. (2008) and Marat et al. (2013) used a face detection module to investigate the role of human faces in the attention selection mechanism. Same rationale has been used for generic objects by Chikkerur et al. (2010) or text (Cerf et al., 2008; Clavelli et al., 2014). Remarkably Torralba (2003) showed how the even gist of the scene, i.e. its semantic category, like "office" or "forest" guides the eye movements.

To be precise, it's worth remarking that the term *saliency* has been historically conceived to describe the topographic representation of the occurrence of bottom-up features. In the visual attention realm when top-down (relevance) and bottom-up (saliency) mechanisms are combined for eye guidance, the resulting map is termed priority map (Egeth and Yantis, 1997). However, in the "modern jargon", despite of this heuristic addition of high-level processing capabilities, these are still referred to as saliency models (Borji and Itti, 2013; Furnari et al., 2014; Bruce et al., 2015; Bylinskii et al., 2015, 2016, 2019). Throughout the following discussions we will deliberately use the term *saliency* or *saliency map* subsuming the latter meaning.

Interestingly enough, mixing bottom-up features and top-down information (mined with the help face or object detectors and scene context) together with the adoption of machine learning techniques to find the best combination of such features, allowed the field of saliency prediction to become a really active subarea of computer vision. Such view of the problem has paved the way to the definition of saliency models in terms of predictors (either, classifiers or regressors), for which a number of learning techniques were readily available. Kienzle et al. (2006) were the first who adopted this approach by learning the discriminant function of patches of the images from eye tracking data using a Support Vector Machine (SVM). A similar approach was then adopted by Judd et al. (2009) who trained a linear SVM from human fixation data using a set of low, middle and high level features to define salient locations. In a similar vein, Yan et al. (2010); Lang et al. (2011); Jiang et al. (2015a) proposed methods relying on sparse representation of "feature words" (atoms) encoded in salient and non-salient dictionaries; these are either learned from local image patches or from eye tracking data of training images. Approaching the problem of saliency prediction from this point of view is appealing, since allows to asses the relevance of visual features in a data driven way by means of optimal predictors (this fact is further exploited in the modern wave of deep saliency models treated below). On the other hand, as remarked by Borji and Itti (2013), this approach makes models "data-dependent, thus influencing fair model comparison, slow, and to some extent, black-box". Indeed is misses the explanatory base as it does not accounts for how attention adapts in humans.

2.2.2 Assessing the performance of saliency models

An issue that straightforwardly raises is how to measure and benchmark the performance of a saliency model accounting for the map $\mathbf{I} \mapsto \mathcal{W}$. The general idea is to measure the capability of the model output, namely the saliency map $\mathcal{W}$, to predict fixations *as if* they were performed. The overall procedure of evaluation is depicted in Figure 2.3.

In a nutshell, eye fixations $\{\mathbf{r}_F^{(s,i)}(1), \mathbf{r}_F^{(s,i)}(2), \cdots\}$ are typically used as to derive the ground-truth. These are collected in an eye-tracking experiment involving $s = 1 \cdots N_S$ subjects on a chosen data set $\{\mathbf{I}^i\}$ of $i = 1 \cdots N_I$ images (or videos); first, raw data is collected from human observers through eye trakers, which record eye gaze position and trajectories. Next, such data is parsed and classified into a sequence of fixatons (scanpaths). Finally, fixation positions are used to build the 2D empirical fixation map. Some metrics use the original binary location map of fixations, say $\mathcal{M}^B$. Alternatively, the discrete fixations can be converted into a continuous distribution, a fixation map (a.k.a *heat map* or *attention map* when fixations are weighted by fixation time), $\mathcal{M}^D$

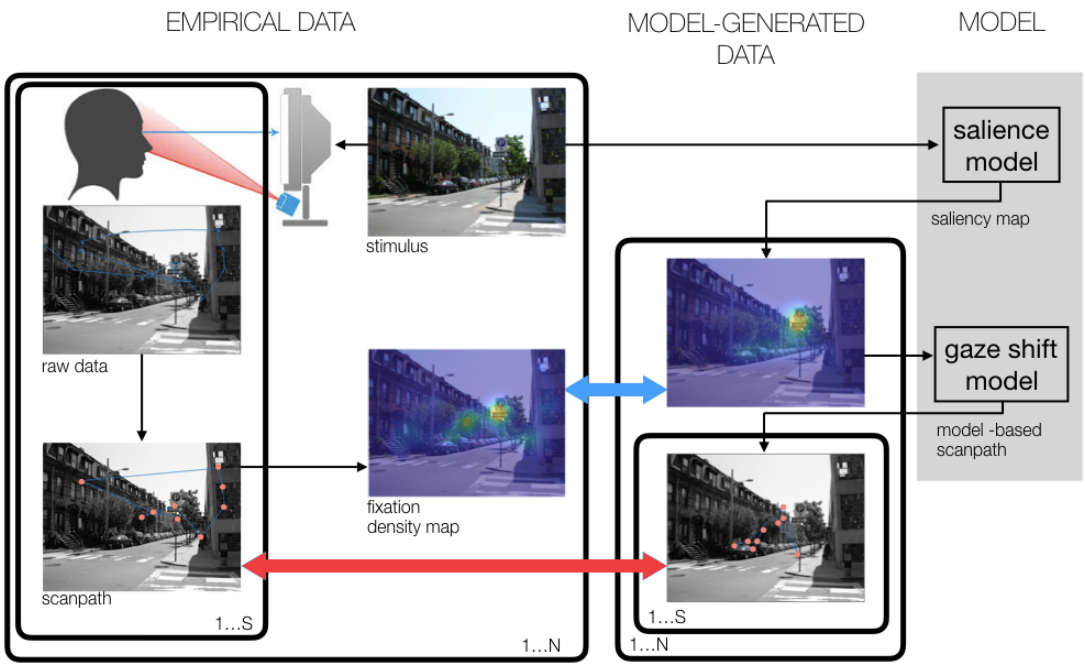


Figure 2.3: *Gaze data recording via eye-tracking and modelling. Given a stimulus (image* $\mathbf{I}$*), the observer's gaze trajectory is sampled and recorded. Raw data are parsed and classified in fixations sequences (scanpaths). Collecting fixations from all subjects the 2D empirical fixation distribution* $\mathcal{M}^D$ *is estimated. On the model side, for the same stimulus a saliency map* $\mathcal{W}$ *is derived; if available, a gaze shift model can be exploited for sampling scanpaths based on* $\mathcal{W}$*. The overall model performance is routinely evaluated by comparing either the model-generated saliency map* $\mathcal{S}$ *with the empirical* $\mathcal{M}^D$ *map (light blue two-head arrow) and/or, albeit less commonly, by confronting the model-generated scanpaths* $\{\widetilde{\mathbf{r}}_F(1), \widetilde{\mathbf{r}}_F(2), \cdots\}$*, with the actual ones* $\{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}$ *(red two-head arrow).*

(Bylinskii et al., 2019). Precisely, for each stimulus $\mathbf{I}^i$ the map

$$\{\mathbf{r}_F^{(s,i)}(1), \mathbf{r}_F^{(s,i)}(2), \cdots\}_{s=1}^{N_S} \mapsto \mathcal{M}^{D(i)}, \tag{2.4}$$

is computed as an empirical fixation density (Kümmerer et al., 2015; Le Meur and Baccino, 2013). Eventually, a metric is evaluated either in the form $\mu(\mathcal{W}, \mathcal{M}^B)$ or $\mu(\mathcal{W}, \mathcal{M}^D)$, the result being a number assessing the similarity or dissimilarity between $\mathcal{W}$, and $\mathcal{M}$. To this end a number of metrics have been proposed (Bylinskii et al., 2019) together with benchmarking datasets (Judd et al., 2012; Borji et al., 2013; Borji and Itti, 2015). This is the typical way of assessing saliency models, in particular is the one adopted in the popular *MIT Saliency Benchmark* (Kümmerer et al., 2018; Bylinskii et al., 2019; Judd et al., 2012).

Less commonly, a gaze shift model can be employed in order perform the mapping $\mathcal{W} \to \{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}$ from the generated saliency map to a new sequence of fixations (model-based scanpath) which is then compared with the real ones (red arrow in Figure 2.3).

The *MIT Saliency Benchmark*, by now, collects the results of more than 90 different models which are compared in terms of 7 different metrics, namely:

- Information Gain (IG) (Kümmerer et al., 2015): is the difference in average log-likelihood between the model's predictions and an image-independent center-bias prior distribution, measured in bits per fixation.
- Area Under the Curve (AUC) (Tatler et al., 2005): The saliency map is treated as a binary classifier to separate positive from negative samples at various thresholds. The true positive (tp) rate is the proportion of saliency map values above threshold at fixation locations. The false positive (fp) rate is the proportion of saliency map values above threshold at all pixels.
- Shuffled Area Under the Curve (sAUC) (Riche et al., 2013): is a version of the Area Under ROC curve measure. The saliency map is treated as a binary classifier to separate positive from negative samples at various thresholds. The true positive (tp) rate is the proportion of saliency map values above threshold at fixation locations. The false positive (fp) rate is the proportion of saliency map values above threshold at pixel locations that are fixated in OTHER IMAGES. In this implementation, all sample values are used as thresholds.
- Normalized Scanpath Saliecy (NSS) (Peters et al., 2005): measures the mean saliency value at fixated locations of the normalized (zero mean, unit variance) saliency map.
- Correlation Coefficient (CC): is the linear correlation coefficient between a model saliency map and an empirical saliency map gained from convolving the fixation locations with a Gaussian kernel.
- Kullback-Leibler Divergence (KLDiv): the Kullback-Leibler divergence normalizes the model's saliency map and an empirical saliency map to be densities by dividing by the sum and then computes the Kullback-Leibler divergence between these two distributions. It is a non-symmetric measure of the information lost when the saliency map is used to estimate the empirical saliency map.

- Similatiy (SIM): this similarity measure is also called histogram intersection and measures the similarity between two different saliency maps when viewed as distributions. It is computed by first normalizing the model's saliency map and an empirical saliency map to be densities by dividing by the sum and then adding the pixelwise minimums of both distributions.

All models are compared in terms of such metrics with a *Gold Standard* and a *Baseline* model.

The *Gold Standard* model is defined as a Gaussian Kernel Density Estimate; there are two versions of it: the crossvalidated performance is the leave-one-subject-out performance where for each subject and image the fixations of all other subject on the same image are used to construct a kernel density estimate that is then evaluated on the remaining subject. The kernel size and the mixture weight of a uniform regularization component are fitted by maximizing the cross-validated log-likelihood of the model. In addition to this crossvalidated version of the model, we also report the performance of a KDE model that uses all fixations on each image with the same parameters as the cross-validated model. One can interpret the cross-validated performance as a lower bound on the explainable performance and the joint performance as an upper bound.

The *Baseline* model is represented by a center bias: is again a Gaussian Kernel Density Estimate. However, unlike the gold standard, it uses the fixations of all other images to predict the fixations on any given image. Kernel size and the mixture weight are again fitted by maximizing the model log-likelihood.

By scrolling on the results of the benchmark two particular facts can be noted: the first one is that many models rank below the center bias baseline model. This list includes the saliency map produced by the Itti et al. (1998) model, thus giving evidence of the importance of top down information. The second thing that stands out, is that the top ranking models mainly belong to the category of *deep saliency models*. Indeed, after a short period of performance saturation around 2010 to 2014, saliency models experienced a sudden burst of performance improvement mainly thanks to the advent of deep learning and the release of large scale crowd sourced data (Jiang et al., 2015b).

2.2.3 The new wave of deep saliency models

Since Krizhevsky et al. (2012) won the ImageNet competition with the Deep Convolutional Neural Network called *AlexNet*, the field of Computer Vision was literally revolutionized. Unavoidably, having become a subfield of Computer Vision, saliency prediction was not immune to such significant shift. The success of CNNs on large scale object recognition tasks, has given rise to a new wave of saliency models performing sensibly better than traditional ones based on hand-crafted features.

The general idea of a deep saliency prediction model is to take advantage of deep models pre-trained on various tasks (object recognition or detection, scene classification) and fine tune them to predict saliency. The turning point in the field was reached when the large scale dataset SALICON (Jiang et al., 2015b) was released, thus enabling a proper fine-tuning even for deeper models. This contributed to a big progress of deep saliency prediction model and several new effective architectures have been proposed (for an exhaustive review of deep saliency models, the reader is redirected to the work of Borji (2019)).

The reason for such effectiveness lies in the possibility of exploiting the concepts learned by the network's convolutional filters for the saliency prediction task. Such concepts may by high level (faces, objects, actions) or low level (orientation, colour, ecc) depending on the position of the filter in the network hierarchy. The role of the deep saliency model is thus to learn to combine the feature activation maps from different layers in order to achieve the *best* mix of bottom-up and top-down features. Indeed, leveraging the end to end nature of such architectures, these concepts can be *learned* rather than hand-crafted. Crucially, the higher level concepts embedded in such deep models are the reason for the big performance gap between early bottom-up approaches and modern ones (Borji, 2019). This fact provides practical evidence of the usefulness of high-level image features for prediction purposes (Bylinskii et al., 2016; Kummerer et al., 2017).

The model proposed by Vig et al. (2014) was the first attempt to apply deep leaning to saliency prediction: their approach basically consists in extracting a large set of features from a pre-trained CNN. The optimal subset of such representations is then found in a data driven way through an optimization procedure. The optimal features are then used as predictors of a linear SVM model which learns the saliecy discriminant function. This work paved the way to the birth of a plenty of new deep saliency models (Borji, 2019).

One of the most known of such models is called *DeepGaze I* by Kümmerer et al. (2014); this model is a pre-trained AlexNet whose outputs of the convolutional layers were used to create and train a linear model to compute image salience. Remarkably, this model reported results which beated the state of the art by a large margin, even if comparing with Vig et al. (2014). More recently Kummerer et al. (2017) released the *DeepGaze II* model; it further explores the unique contributions between low-level and high-level features towards fixation prediction and exhibits better performances than his ancestor *DeepGaze I*. It is built upon the more recent (and deeper) VGG-19 architecture (Simonyan and Zisserman, 2014), trained to identify objects in images. In particular, the activations of a subset of the pre-trained VGG feature maps for a given image are passed to a second neural network (the readout network) consisting of four layers of 1×1 convolutions. The parameters of VGG are held fixed through training, thus only the readout network learns about saliency prediction. This results in a final saliency map, which is then blurred, combined with a center bias and converted into a probability distribution by means of a softmax function. The computation of the center bias, acting as a prior, is carried out by averaging all the fixations collected from real observers on the training dataset (Kummerer et al., 2017).

2.2.4 A criticism to saliency maps

Crucially, saliency maps do not account for temporal dynamics. They are by and large spatially evaluated across all fixations, precisely by comparing to maps $\mathcal{M}^B$, or $\mathcal{M}^D$ derived from fixations accumulated in time after the stimulus onset until the end of the trial (Eq. 2.4).

As a matter of fact, surmising that $\mathcal{W}$ is predictive of human fixations does not entail an actual mechanism of fixation generation, $\mathcal{W}_i \mapsto \{\widetilde{\mathbf{r}}_F^{(s,i)}(1), \widetilde{\mathbf{r}}_F^{(s,i)}(2), \cdots\}$ to be compared against actual fixation sequences $\{\mathbf{r}_F^{(s,i)}(1), \mathbf{r}_F^{(s,i)}(2), \cdots\}$. The assessment of

the predictive capability of a model is just to be understood as the indirect measurement of any metric μ as introduced above. When using the mapping of Eq. 2.4, it is implicitly assumed that fixations, once collected, are exchangeable with respect to time ordering $\{1, \cdots, n\}$, namely

$$\{\mathbf{r}_F^{(s,i)}(1), \mathbf{r}_F^{(s,i)}(2), \cdots \mathbf{r}_F^{(s,i)}(n)\} = \{\mathbf{r}_F^{(s,i)}(\pi(1)), \mathbf{r}_F^{(s,i)}(\pi(2)), \cdots, \mathbf{r}_F^{(s,i)}(\pi(n))\}, \quad (2.5)$$

$\forall \pi \in \Pi(n)$ where $\Pi(n)$ is the group of permutations of $\{1, \cdots, n\}$. This assumption implies that any dynamical law $\widetilde{\mathbf{r}}_F^{(s,i)}(t) = f(\widetilde{\mathbf{r}}_F^{(s,i)}(t-1), \mathcal{W}_i)$ that takes as input the perceptual representation of the i-th image and the previous fixation location (as a system state) and returns the next location of fixation as its output is dismissed. However, dynamics is important in many respects. For instance, there is evidence for the existence of systematic tendencies in oculomotor control (Tatler and Vincent, 2009): eyes are not equally likely to move in any direction. Yet, apart from the well known center bias (Tatler, 2007), motor biases can be actually taken into account only when scanpath generation is performed.

In such perspective, Le Meur and Coutrot (2016) have proposed saccadic models as a new framework to predict visual scanpaths of observers while they freely watch static images. In such models the visual fixations are inferred from bottom-up saliency and oculomotor biases (captured as saccade amplitudes and saccade orientations) that are modeled using eye tracking data. Performance of these models can be evaluated either by directly comparing the generated scanpaths to human scanpaths or by computing new saliency maps, in the shape of densities from model generated fixations (red arrow in Figure 2.3). There is a limited number of saccadic models available, see Le Meur and Coutrot (2016) for a comprehensive review; generalisation to dynamic scenes have been presented for instance in Boccignone and Ferraro (2014); Napoletano et al. (2015). A remarkable result obtained by saccadic models is that by using simulated fixations $\{\widetilde{\mathbf{r}}_F^{(s,i)}(1), \widetilde{\mathbf{r}}_F^{(s,i)}(2), \cdots\}$ to generate a model-based fixation map, the latter has higher predictive performance than the raw salience map $\mathcal{W}$, in terms of similarity/dissimilarity μ with respect to human fixation maps. Beyond the improvement, it is worth noting that even in this case the model-generated attention map is eventually obtained *a posteriori*, as a 2-D spatial map of accumulated fixations. Such problem is somehow attenuated when dynamic stimuli (videos) are taken into account, though, the temporal unfolding as learned in a data-driven way presents complex albeit structured temporal patterns (Boccignone et al., 2019b; Coutrot and Guyader, 2014b), that deserve being taken into consideration.

2.3 The unfolding of visual attention (and gaze shifts)

Going back to the broader description of a computational model of attentive eye guidance given in Eq. 2.1, if the fist step (Eq. 2.2) has received a lot of attention as witnessed by the huge literature and the raise of benchmarking competitions, the second one (Eq. 2.3), has been much less noticed. This is surprising, given that the most cited work in the field (Itti et al., 1998), explicitly addresses the problem of *How* to look at a picture rather than just *Where*, albeit by means of a simple WTA procedure.

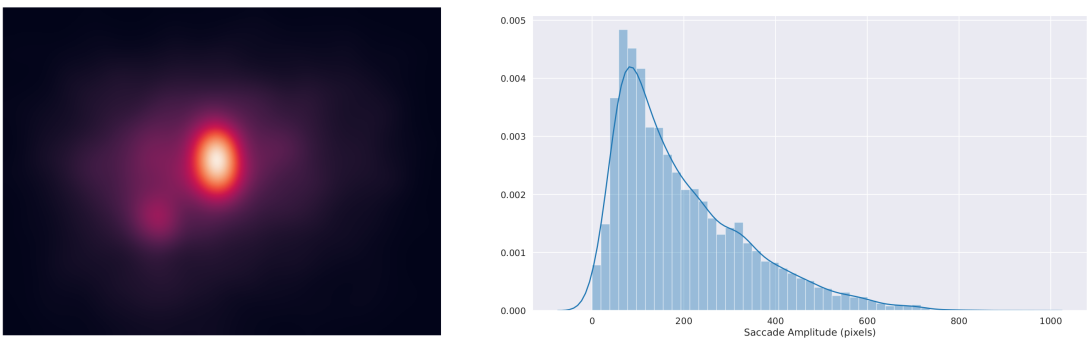
Answering the *How* question when dealing with visual attention brings with it some hurdles; indeed this single statement hides a number of other related questions: *How*

much time do I need to spend looking at a certain portion of the stimuli?, *How do I select which part of the scene is "interesting" at a given time?* and *How do I decide where to look next?*.

Giving an exact answer to each of these question is tricky, mainly due to the fact that the attentive process is subject to a certain amount of randomness: as a matter of fact two people looking at the same exact stimuli will surely produce different gaze position sequences; indeed the same fact happens even in case the same person looks at the same scene twice. This is probably likely to be originated from endogenous stochastic variations that affect each stage between a sensory event and the motor response: sensing, information processing, movement planning and executing (van Beers, 2007). Nonetheless, the study of eye movements revealed many *systematic tendencies* and common biases as well as a structured dynamics (Tatler and Vincent, 2008; Schütt et al., 2019). As a matter of fact, when dealing with gaze shifts a number of cues should be taken into account, namely *systematic tendencies*, *variability* and *dynamics*.

2.3.1 Systematic tendencies and biases

It has been widely demonstrated that regardless of the visual stimuli, the gaze behavior exhibits some systematic tendencies and biases; these can be thought of as regularities that persist across all instances of, and manipulations to, behavioural tasks (Tatler and Vincent, 2008). There exists a good number of such tendencies that characterize gaze behavior. One remarkable example is the amplitude distribution of saccades and microsaccades that typically exhibit a positively skewed, long-tailed shape (Tatler et al., 2011; Dorr et al., 2010; Tatler and Vincent, 2008). This is shown in Figure 2.4b, where the empirical distribution of saccades amplitude collected on the *MIT1003* dataset (Judd et al., 2009) is depicted. Other paradigmatic examples of systematic tendencies in scene viewing are: initiating saccades in the horizontal and vertical directions more frequently than in oblique directions; small amplitude saccades tending to be followed by long amplitude ones and vice versa (Tatler and Vincent, 2008) or the attraction of the observer towards the center of the image (*Center Bias*). The latter is depicted in Figure 2.4a which shows the heatmap of all the fixation recorded from human subjects on the different images of the *MIT1003* dataset. Notably, most of the fixation appear to be concentrated in the center of the image.



(a) *Heatmap of all the fixations*

(b) *Saccades amplitude distribution*

Figure 2.4: *Systematic tendencies on the MIT1003 dataset.*

This tendency is caused by a number of possible factors: displacement bias of an image content (known as photographer bias), motor bias (related to the experiment protocol) as well as physical preferences in orbital position (Tseng et al., 2009; Rothkegel et al., 2017). In a remarkable study, Tatler and Vincent (2009) provided striking evidence that a model based solely on these biases and therefore blind to current visual information can outperform salience-based approaches.

2.3.2 Variability

When looking at a scene, even though the attentional attractors may be the same and despite the systematic tendencies, different subjects will look in different ways. In other words, there is a small probability that two observers will fixate exactly the same location at exactly the same time. This effect is observable either in free viewing conditions or task specific ones (in the former case the effect being more marked). Such important amount of variability can be noticed even when semantically rich objects like faces are present; this fact can be appreciated by looking at Figure 2.5. It can be noted as the 5 different subjects (Figure 2.5(b) to Figure 2.5(f)), although spotting more or less the same regions of the image (face, hands), have different exploration strategies.

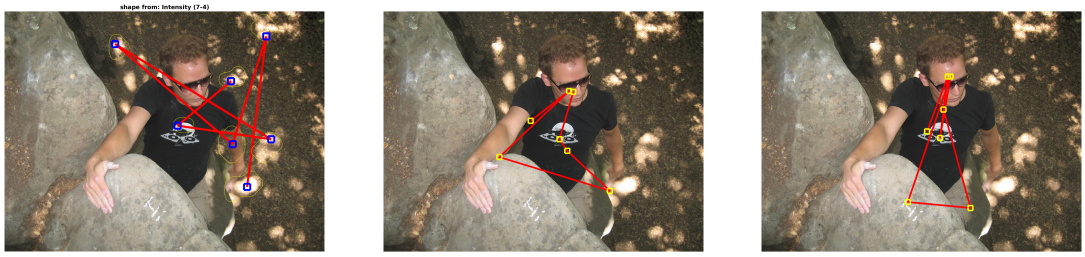
Notably, consistency in fixation locations selected by observers decreases over the course of the first few fixations after stimulus onset (Tatler et al., 2011) and can become idiosyncratic. Nonetheless, variability is also exhibited by the same subject along different trials on equal stimuli.

In the literature, few works have addressed the problem of variability. The WTA approach proposed by Itti et al. (1998), or variants such as the selection of the proto-object with the highest attentional weight (Wischnewski et al., 2010) are themselves deterministic procedures. Even when probabilistic frameworks are used to infer where to look next, the final decision is often taken via the maximum a posteriori (MAP) criterion which again is a deterministic procedure (Elazary and Itti, 2010; Boccignone, 2008; Najemnik and Geisler, 2005; Chernyak and Stark, 2001).

Figure 2.5(a) depicts the typical output of the Itti's model (Itti et al., 1998), i.e. the sequence of fixations produced by the WTA procedure. Setting aside the fact that the lack of top down information prevents the model to correctly select the high level cues of the scene (face, hands ecc.), the important thing to note here is that the WTA approach will produce the same output if ran multiple times. In other words, given the stimulus $\mathbf{I}$, the mapping $\mathcal{W} \rightarrow \{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}$ is a deterministic function. Conversely, the sequence of eye movements on a given stimuli should be conceived, in a more realistic way, as a realization of a stochastic process.

2.3.3 Dynamics

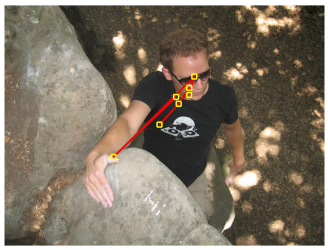
In Section 2.2.4 we made a criticism on the use of saliency maps as a *proxy* for attention deployment. Indeed in a *plain* saliency map the temporal information of each of the fixations has been squeezed and the map shows the likelihood of fixating a particular region of the stimuli as "freezed" at the end of the viewing process (i.e, after having collected all fixations on stimulus along an eye-tracking session). As a consequence saliency maps are not able to describe the dynamic nature of visual attention. It's worth saying that such pitfall is somewhat mitigated when considering dynamic



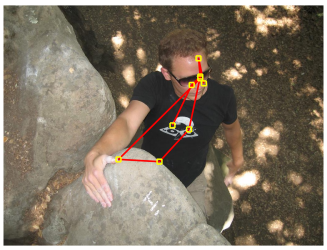
(a) *Deterministic WTA mechanism*

(b) *Subject 1*

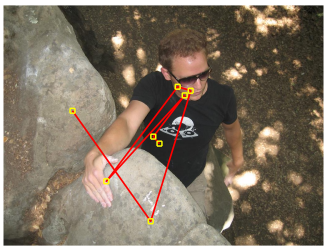
(c) *Subject 2*



(d) *Subject 3*



(e) *Subject 4*



(f) *Subject 5*

Figure 2.5: *(a) Sequence of fixatons produced by the WTA procedure of Itti et al. (1998). (b)-(f) Sequence of fixations for 5 different subjects from the MIT1003 dataset.*

stimuli (videos), but is widely overlooked on static ones (images). Indeed, even static stimuli yield an attention allocation with it's own dynamics. This very fact has been demonstrated by Schütt et al. (2019); in this recent work, the authors have for the first time, considered the temporal evolution of the fixation density in the free viewing of static scenes. They provide evidence for a fixation dynamics which unfolds into three phases:

1. An initial orienting response towards the image center;

2. A brief exploration, which is characterized by a gradual broadening of the fixation density, the observers looking at all parts of the image they are interested in;

3. A final equilibrium state, in which the fixation density has converged, and subjects preferentially return to the same fixation locations they visited during the main exploration.

Beyond the theoretical insights offered by their analyses, by monitoring the performance of the empirical fixation density over time, they also pave the way to a more subtle and principled approach to unveil the actual predictive performance of saliency models (Schütt et al., 2019). It thus may be interesting to understand to what extent the inclusion of the dynamics of gaze shifts when free viewing static images, brings some benefits. At least from the theoretical point of view, neglecting such information, by considering a static saliency map as *per se* predictive of overt attention may lead to sub-optimal results.

2.3.4 Gaze shift models

The operational description of all the aspects described thus far is accomplished via *Gaze Shift* or *Saccadic* models. These are procedures that ingest a perceptual representation of the stimuli $\mathcal{W}$ (whatever it may be) and produce a sequence of fixation locations. In what follows, we give a brief overview of some of the saccadic models that have been proposed in the literature, highlighting the distinction between the models conceived to deal with either static or dynamic stimuli. Unfortunately, only a handful of models have been proposed for predicting gaze shift dynamics, and most of these are conceived for processing static image input.

Gaze shift models on static stimuli

When dealing with static stimuli (images), the most celebrated example is of course the Itti et al. (1998) model, the pioneering work of the field; as previously pointed out, however, this model lacks some of the critical aspects of of gaze shifts, namely variability and the modelling of systematic tendencies.

Some of these aspects where lately addressed, for instance by Le Meur and Liu (2015) who proposed a model to predict observers' scan paths on static images relying on bottom-up saliency; most importantly, the oculomotor biases are taken into account by sampling from empirical distributions of saccades amplitude and orientation computed on publicly available datasets.

Wang et al. (2011) proposed a computational model of scanpath prediction based on the principle of information maximization. The model integrates three related factors as driven forces to guide eye movements sequentially - reference sensory responses, fovea-periphery resolution discrepancy, and visual working memory.

Sun et al. (2014) presented a statistical framework for modelling both saccadic eye movements and visual saliency which are modelled based on super-Gaussian component (SGC) analysis.

Wloka et al. (2018) proposed STAR-FC, a model for saccades generation on static stimuli based on the integration of high-level and object-based saliency and peripheral lower-level feature-based saliency

More recently, solutions relying on deep networks have been proposed; Assens et al. (2018a) proposed PathGAN, a deep neural network for visual scanpath prediction trained on adversarial examples. The same authors presented a deep model called SaltiNet (Assens et al., 2018b) to predict scanpaths and saliency on 360-degree images.

Xia et al. (2019) address the problem of saccadic scanpath prediction by introducing an iterative representation learning framework in which eye movements are the outcome of the current representation which is then updated based to the gaze shift. The fixation selection relies on a perceptual residual which is computed by means of a an auto-encoder network. Xia and Quan (2020) used a similar iterative representation model to predict human scanpaths of web pages. Bao and Chen (2020a) proposed a deep convolutional saccadic model which simultaneously predicts the foveal saliency maps and fixation durations, both aspects are handled by convolutional neural networks (CNNs). Sun et al. (2019) proposed a recurrent mixture density network based framework to predict human-like scanpaths on static images; the model predicts both the sequence of fixations and their duration relying on both bottom up saliency and seman-

tic features extracted by convolutional neural networks.

Of key interest for the forthcoming sections of the present chapter is the model proposed by Boccignone and Ferraro (2004), named *Constrained Lèvy Exploration* (CLE); briefly, the CLE considers the gaze motion as given by the stochastic dynamics of a Lévy forager moving under the influence of an external force (which, in turn, depends on a salience or attention potential field). Namely, at time t the transition from the current position $\mathbf{r}(t)$ to a new position $\mathbf{r}_{new}(t)$, $\mathbf{r}(t) \rightarrow \mathbf{r}_{new}(t)$, is given by:

$$\mathbf{r}_{new}(t) = \mathbf{r}(t) + \mathbf{g}(\mathcal{W}(\mathbf{r}(t))) + \boldsymbol{\eta}. \tag{2.6}$$

The trajectory of the variable $\mathbf{r}$ is determined by a deterministic part $\mathbf{g}$, the drift - relying upon salience or fixation density -, and a stochastic part $\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a random vector sampled from a heavy-tailed distribution, accounting for motor biases.

The Lévy forager's dynamics formalised in Eq.2.6 can be written:

$$\mathbf{r}_{new}(t) = \mathbf{r}(t) - \nabla V + \boldsymbol{\eta}, \tag{2.7}$$

so that the new gaze position is determined by: a) the gradient of V, the external force field shaped by the perceptual landscape, $V(\cdot, t)$ being defined as the time varying scalar field

$$V(x, y, t) = \exp(-\tau_V \mathcal{W}(x, y, t)), \tag{2.8}$$

b) the stochastic vector $\boldsymbol{\eta}$ with components

$$\eta_x = l\cos(\theta), \qquad \eta_y = l\sin(\theta), \tag{2.9}$$

where the angle θ represents the flight direction and l is the jump length. Direction and length are sampled from the uniform and α-stable distribution, respectively:

$$\theta \sim Unif(0, 2\pi), \tag{2.10}$$

$$l \sim \varphi(\mathcal{W}) f(l; \alpha, \beta, \gamma, \delta). \tag{2.11}$$

Along the extensive stage, θ and l summarise the internal action choice of the forager and the function $\varphi(\mathcal{W})$ modifies the pure Levy flight, since the probability to move from one site to the next site depends on the "strength" of a bond

$$\varphi(\mathcal{W}) = \frac{\exp(-\beta_P(\mathcal{W}(\mathbf{r}(t)) - \mathcal{W}(\mathbf{r}_{new}(t))))}{\sum_{\mathbf{r}'_{new}} \exp(-\beta_P(s(\mathbf{r}(t)) - \mathcal{W}(\mathbf{r}'_{new}(t))))} \tag{2.12}$$

that exists between them. The shift proposal is weighed up according to an accept/reject Metropolis rule that depends on the perceptual gain $\Delta\mathcal{W}$ and on "temperature" T. The values of T determine the amount of randomness in scanpath generation. If no suitable shift $\mathbf{r}(t)_{new}$ has been selected, the current fixation point $\mathbf{r}(t)$ is retained.

Gaze shift models on dynamic stimuli

For what concerns gazes shift models conceived to deal with dynamic stimuli, only few models have been proposed:

Boccignone and Ferraro (2014) proposed *Ecological Sampling*, a stochastic model of eye guidance on videos, which assumes the gaze sequence to be generated by a stochastic process. The gaze shift dynamics is implemented in terms of a stochastic differential equation driven by α-stable noise, and grounds its motivation in the Lévy flight approaches to foraging displacements (Viswanathan et al., 2011; Wosniack et al., 2017). The perceptual representation is formalised in terms of proto-objects, i.e. units of visual information that can be accessed by selective attention and subsequently validated as actual objects. The eye guidance strategy consists of choosing where to look next by sampling the appropriate motor behaviour (i.e., the action to be taken: fixating, pursuing or saccading), conditioned on the perceived world and on previous actions. The overall control strategy is based on a complexity measure of the perceived time-varying scene, while the behavioural state (i.e., the action to be taken: fixating, pursuing or saccading) is obtained by a composite sampling strategy which depends on the complexity of the perceived scene at a given time. Complexity is computed from interest points that are stochastically sampled from the proto-object representation.

More recently, Zanca et al. (2019) proposed *G-Eymol*; the model generates gaze trajectories via differential equations of motion derived through variational laws somehow related to mechanics. The focus of attention is subject to a gravitational field. The distributed virtual mass that drives eye movements is associated with the presence of details and motion in the video. The inhibition of return (IOR, Itti and Koch (2001)) mechanism is employed to avoid the model being stuck in the same portions of the visual landscape. Unlike most current models, the proposed approach does not estimate directly the saliency map, but the prediction of eye movements allows to integrate over time the positions of interest. The process of inhibition-of-return is also supported in the same dynamic model with the purpose of simulating fixations and saccades. The differential equations of motion of the proposed model are numerically integrated to simulate scanpaths on both images and videos.

The virtual masses are proportional to the amount of details and motion of the scene, defined as the magnitude of the gradient and the magnitude of the optical flow, respectively. A so defined model, clearly relies on a purely bottom-up approach. However, authors suggest that top-down information can considered by defining object-based gravitational attractors. The original implementation relies on the Haar cascade face detection (Viola and Jones, 2004), that allows faces as additional masses. The *G-Eymol* equation of motion are deterministic. However, the stochasticity requested to sample different scan paths mimicking different observers can be achieved by perturbing the initial conditions of the equations.

2.3.5 Evaluation of gaze shift models

Similarly to the saliency models, the assessment of gaze shift models raises the problem of defining a performance metric able to capture all the hurdles carried by a fixation sequence. Unlike classic work on saliency estimation, where standard metrics are avail-

able and widely adopted, here there is a lack of consensus about the most appropriate evaluation metrics to be used (Le Meur and Baccino, 2013; Anderson et al., 2015).

The evaluation procedure of gaze shift models can be visually described by following the red arrow in Figure 2.6. In particular, given the real scanpath recorded from the s-th subject $\mathcal{R}_s$ and the u-th fixation sequence generated from the gaze shift model to be evaluated $\mathcal{R}_u$, a scanpath metric is the function $\mu(\mathcal{R}_s, \mathcal{R}_u)$ returning a number (or a vector of values) representing the degree of similarity/dissimilarity between the two fixation sequences.

In recent years, a number of measures have been proposed, each one able to deal with specific aspects of scan path similarity; as a consequence, the choice of the appropriate scanpath metric depends on the particular feature that one wants to measure. Some of the most widely used metrics and their qualitative behaviour are briefly recapped below, but for an in-depth review and discussion see Anderson et al. (2015).

ScanMatch

One of the most successful ways of comparing scanpaths is based on the *string edit distance*, normally used to compare sequences of characters. In particular, a set of edits (insertions, deletions, substitutions) are preformed to transform one string into the other. The similarity between the two sequences is given by the number of editing steps required for the transformation. This method has been adopted for the comparison of scanpaths (Brandt and Stark, 1997). In order to do so, the image is divided in cells to which is assigned a unique character. Fixation sequences can thus be treated as sequences of characters. This method has been later refined by Cristino et al. (2010) who proposed the *ScanMatch* metric.

ScanMatch is based on the Needleman–Wunsch algorithm used in bioinformatics to compare DNA sequences (Cristino et al., 2010). The two fixation sequences are first spatially and temporally binned, then re-coded in order to obtain sequences of letters that represent the spatial (position), temporal (duration) and order information. The obtained letter sequences are then compared by maximizing the similarity score computed from a substitution matrix that provides the score for all letter pair substitutions and a penalty gap. The algorithm returns for each pair of scan paths a score in the range $[0, 1]$. The main advantage of ScanMatch is that it can take into account spatial, temporal, and sequential similarity between scanpaths, thus giving an overall summary of the resemblance between two fixation sequences; on the other hand ScanMatch is not able to provide a description of the performance w.r.t. the different dimensions of gaze dynamics. Moreover it suffers from the quantization issues inherent to the spatial and temporal binning process.

MultiMatch

The MultiMatch metric (Jarodzka et al., 2010; Dewhurst et al., 2012) is a multi-dimensional, vector-based method to measure scanpath similarity. The algorithm ingests the two sequences to be compared, which may differ in length; both scanpaths are then *simplified*, i.e. successive fixations are combined if they are within a given distance or within a given directional threshold of each other. Subsequently, scanpaths are aligned based on their shape using a dynamic programming approach. Finally, the method returns as

output a 5-dimensional vector, each dimension describing the degree of similarity with respect to different aspects of the scanpaths, namely:

- *Shape*: Is computed as the vector difference between aligned saccade pairs, normalized by the screen diagonal. It is sensitive to spatial differences in fixation position and measures the overall similarity in shape between the two fixation sequences.
- *Length:* Is computes as the absolute difference in length between endpoints of aligned saccade vectors, normalized by the screed diagonal. This measure is sensitive specifically for the saccades amplitude and discards information related to other aspects like direction, position or duration of fixations.
- *Direction:* Is the angular distance between aligned saccade vectors normalized by π. This dimension gives precise insights on the similarity between saccades direction, bt not on any other feature.
- *Position:* Is the Euclidean distance between aligned fixations, normalized by the screen diagonal. This measure is also sensitive to saccades amplitude and direction.
- *Duration:* Is the absolute difference in fixation duration between aligned fixations, normalized by the maximum duration.

Each dimension is normalised in order to have values in the range $[0, 1]$, higher values meaning higher similarity $(1 - distance)$.

The main advantage of the MultiMatch method is that it provides several measures to assess scanpath similarity, each measure capturing a unique aspect of scanpath resemblance. On the other hand, the scanpath simplification procedure makes unclear how sensible is the metric with respect to variations (Anderson et al., 2015).

Recurrence Quantification Analysis (RQA)

Recurrence Quantification Analysis (RQA) is typically exploited to describe complex dynamical systems. Recently (Anderson et al., 2013) it has been adopted to quantify the similarity of a pair of fixation sequences by relying on a series of measures that are found to be useful for characterizing cross-recurrent patterns (Anderson et al., 2015). Given two fixation sequences, $\mathcal{R}_1$ and $\mathcal{R}_2$, RQA calculates the cross-recurrence for each fixation of two scanpaths $\mathcal{R}_{1,i}$, $\mathcal{R}_{2,j}$ (for scanpaths of different lengths, the longer one is trimmed to the shortest), resulting in the construction of the so-called recurrence plot: two fixations are cross-recurrent if they are close together in terms of their Euclidean distance. Cross-recurrence c_{ij} between the i-th and j-th fixations of the two scanpaths to be compared, can thus be defined as:

$$c_{ij} = \begin{cases} 1, & d\left(\mathcal{R}_{1,i}, \mathcal{R}_{2,j}\right) \leq \rho \\ 0, & \text{otherwise} \end{cases} \tag{2.13}$$

Where d is the distance metric (typically Euclidean distance) and ρ is a given radius. Cross-recurrence can be represented in a cross-recurrence diagram: if two fixations are

cross-recurrent, then ($c_{ij} = 1$), then a dot is plotted in position i, j. The measures that have been found useful for characterizing cross-recurrent patterns are described in the following:

Cross-recurrence Represents the percentage of fixations that match between the two fixation sequences. Intuitively, this measure gives the degree of similarity in fixation position between to scanpaths. The more similar the two fixation sequences, the higher the number of ones in the cross-recurrence diagram. It is invariant to the order of fixation.

Determinism Represents the percentage of fixation trajectories common to both scanpaths. In other words, it quantifies the overlap of a specific sequence of fixations, preserving their sequential information. Although two scanpaths may be quite dissimilar in their overall shape or fixation positions, this measure may show whether certain smaller sequences of those scanpaths may be shared.

Laminarity Measures how much the two fixation sequences cluster together. It represents locations that were fixated in detail in one of the fixation sequences, but only fixated briefly in the other fixation sequence.

Center of recurrence mass Center of recurrent mass (CORM) indicates the dominant lag of cross-recurrences. Small CORM values indicate that the same fixations in both fixation sequences tend to occur close in time, whereas large CORM values indicate that cross-recurrences tend to occur with either a large positive or negative lag.

2.4 Evidence of attention dynamics through gaze shift models

The previous section highlighted the fact that according to recent studies (Schütt et al., 2019), attention allocation dynamics on static stimuli may exhibit a defined structure. Based on such claim, we propose a complementary analysis that relies on model-generated scanpaths, i.e. actual prediction. More precisely, we ask the following: do model-generated scanpaths differ from human scanpaths in the free viewing of static scenes when 1) the scanpath is generated by taking into account the time varying evolution of attention (cfr. Section 2.3.3) as opposed to when 2) the scanpath is generated by only taking into account the final fixation density?

The importance of modelling dynamics in gaze deployment can be proved by a straightforward experiment (Boccignone et al., 2019a): a time-varying fixation density is used as the attention map that moment-to-moment feeds the gaze shift dynamics. In other words rather than freezing the map to final fixations we compute different empirical fixation maps (computed for the specific scene from other observers' behaviour) at different time steps, in order to take into account the unfolding of attention dynamics as described in Section 2.3.3. Using the empirical fixation map rather then a "predicted" one, allows us to assess differences rising at the oculomotor behavior while being free from any saliency model specific assumption. In brief we do the following:

Step 1 Compute three different empirical fixation density maps $\mathcal{M}_k^{D(i)}$ accounting for

phases $k = 1, 2, 3$ above (cfr Section 2.3.3), by aggregating all the human fixations performed in the corresponding time window (first 3 images of Figure 2.6):

$$\{\mathbf{r}_F^{(s,i)}(m_{k-1}+1), \cdots, \mathbf{r}_F^{(s,i)}(m_k)\}_{s=1}^{N_S} \mapsto \mathcal{M}_k^{D(i)}, \quad k = 1, 2, 3. \tag{2.14}$$

Step 2. Generate "subject" fixations depending on the three-phase unfolding defined above, by relying on a saccadic model $\mathbf{r}_F^{(s,i)}(n) = f(\mathbf{r}_F^{(s,i)}(n-1), \mathcal{W}(k)_i)$:

$$\mathcal{M}_k^{D(i)} \mapsto \{\widetilde{\mathbf{r}}_F^{(s,i)}(m_{k-1}+1), \cdots, \widetilde{\mathbf{r}}_F^{(s,i)}(m_k)\} = \mathcal{R}t_k^{(s,i)}, \quad k = 1, 2, 3 \tag{2.15}$$

with $\mathcal{W}(k)_i = \mathcal{M}_k^{D(i)}$ being the phase-dependent perceptual representation of image i, so to obtain the "time-aware" scanpath $\mathcal{R}t^{(s,i)} = \{\mathcal{R}t_1^{(s,i)}, \mathcal{R}t_2^{(s,i)}, \mathcal{R}t_3^{(s,i)}\}$.

For comparison purposes, in the same way, but only by relying on the overall final fixation map $\mathcal{M}^{D(i)}$, we perform the mapping $\mathcal{M}^{D(i)} \mapsto \mathcal{R}s^{(s,i)}$, which represents the typical output of a saccadic model.

It should be intuitively apparent that the evolution of the empirical fixation density $\mathcal{M}_t^{D(i)}$ within the time interval $[t_0, T]$ from the onset of the stimulus i up to time T, provides a source of information which is richer than that derived by simply considering its cumulative distribution function $\int_{t_0}^{T} \mathcal{M}_t^{D(i)} dt$. Yet, this very fact is by and large neglected in the saliency modelling practice. A so defined experiment would have the virtue of relying on a generative approach taking into account the empirical findings presented in Schütt et al. (2019), albeit needing a suitable operational definition of the aforementioned three phases. This latter aspect will be covered in the next section.

Simulation The simulation procedure goes as follows: we generate four different attention maps for each image $\mathbf{I}^i$ of the dataset presented in Judd et al. (2009). Three of these are the temporal density fixation maps $\mathcal{M}_1^{D(i)}, \mathcal{M}_2^{D(i)}, \mathcal{M}_3^{D(i)}$, with $t_{m_1} = 1$, $t_{m_2} = 2$ and $t_{m_3} = 3$ seconds (Eq. 2.14); the fourth is the classic, cumulative $\mathcal{M}^{D(i)}$ map. Fig. 2.6 shows one example.

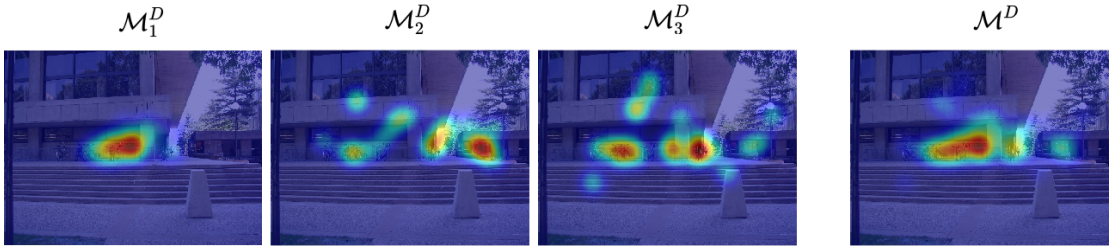


Figure 2.6: *Example of different fixation density maps for a specific image. From left to right: the three temporal distribution maps obtained from fixations collected at seconds* 1*,* 2 *and* 3*, respectively, overlapped on the original stimulus; the standard fixation map resulting from the aggregation of all fixations available at the end of the eye-tracking procedure. The latter map is the one typically exploited in saliency modelling and benchmarking.*

These were used to support the generation of $N_S = 15$ scanpaths for both the temporal (Eq. 2.15) and the classic approach, collected into the sets $\mathcal{R}t^{(i)}$ and $\mathcal{R}s^{(i)}$,

respectively. To such end we exploit the Constrained Levy Exploration (CLE) (Boccignone and Ferraro, 2004)) saccadic model that has been widely used for evaluation purposes (Le Meur and Coutrot, 2016; Xia et al., 2019).

Figure 2.7 shows CLE generated scanpaths, compared against the actual set of human scanpaths $\mathcal{R}^{(i)} = \{\mathbf{r}_F^{(i)}(1), \cdots, \mathbf{r}_F^{(i)}(m_3)\}$.

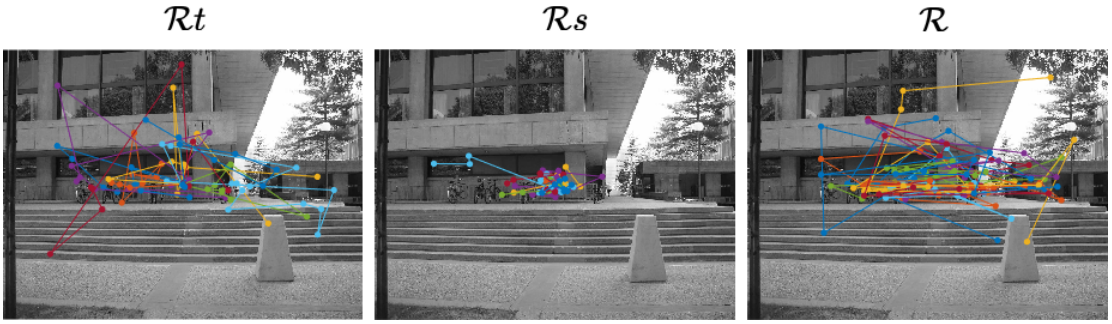


Figure 2.7: *Scanpaths for the image in Fig. 2.6. Left to right: 15 model-generated scanpaths, via Eq. 2.6 from the temporally unfolded fixation maps, 15 model-generated scanpaths from the standard fixation map, 15 scanpaths from actual human fixation sequences (ground-truth). Different colours encode different "observers"(artificial or human).*

The example shows at a glance that when attention deployment is unfolded in time, the predicted scanpaths more faithfully capture the dynamics of actual scanpaths than the dynamics of those generated via the "freezed" map. To quantitatively support such insight, the quality of $\mathcal{R}t^{(i)}$ and $\mathcal{R}s^{(i)}$ has been evaluated on each image i of the dataset by adopting metrics based on the ScanMatch (Cristino et al., 2010) and the recurrence quantification analysis (RQA) (Anderson et al., 2013))[1].

Results All the generated scanpaths belonging to $\mathcal{R}t$ and $\mathcal{R}s$ have been evaluated against the human ones $\mathcal{R}$ for each image. Table 2.1 reports the average values over all the "observers" related to the same images in the dataset. To quantify the intra-human similarity, an additional measure resulting from the comparison of $\mathcal{R}$ with itself is provided. It can be noticed that the temporal approach outperforms the static one in all the three adopted metrics, thus giving evidence of the benefits that may come from modeling attention dynamics.

It's worth noticing how the Determinism score for the $\mathcal{R}t$ is actually greater for the simulation than for the inter-observer comparison; this probably means that, for this metric, the model simulated scanpaths lack some variability if compared with the real ones. By recalling the definition of Determinism, this may suggest that the model generated scanpaths present some sub-sequences that are highly similar to those of real subjects. On the other hand, when comparing real observers, such sub-sequences may exhibit less fixed patterns that are averaged out in the empirical fixation map computation.

[1] An implementation is provided at `https://github.com/phuselab/RQAscanpath`

	ScanMatch	Determinism	CORM
$\mathcal{R}s$ **vs.** $\mathcal{R}$	0.39 (0.08)	58.08 (11.18)	19.95 (5.90)
$\mathcal{R}t$ **vs.** $\mathcal{R}$	**0.43** (0.05)	**61.65** (8.51)	**15.26** (3.58)
$\mathcal{R}$ **vs.** $\mathcal{R}$	0.49 (0.05)	59.61 (7.71)	10.0 (2.09)

Table 2.1: *Average values (standard deviations) of the considered metrics evaluated over all the artificial and human "observers" related to the same images in the dataset.*

2.4.1 A model for time-aware scanpath generation

So far we gave evidence of the existence of a temporal dynamics which affects the gaze deployment to static stimuli by means of a straightforward procedure relying on a *time-aware* empirical fixation density map. The main goal of this section is thus to outline a model to substantiate such results. In essence, rather than relying on empirical fixation maps, we propose a *time-aware* computational model to predict gaze shifts on new images.

In brief, the scheme we propose consists of a three-stage processing where the dynamics described by Schütt et al. (2019) basically relies on:

1. a center-bias model for initial focusing;
2. a context/layout model accounting for the broad exploration to get the gist of the scene;
3. an object-based model, to scrutinize objects that are likely to be located in such context.

The output of each model is a specific map, guiding, at a that specific stage, the sequential sampling of a partial scanpath via the gaze shift model. The three-stage model is outlined at a glance in Fig. 2.8. The "time-aware" scanpath $\mathcal{R}t^{(s,i)} = \{\mathcal{R}t_1^{(s,i)}, \mathcal{R}t_2^{(s,i)}, \mathcal{R}t_3^{(s,i)}\}$, for each "artificial observer" s viewing the i-th stimulus, is obtained from the three partial scanpaths. These are sampled by relying on the three maps computed via the center bias, context and object models, respectively. Each model m is activated at a delay time D_m, while inhibiting the output of model $m-1$, so that the gaze model operates sequentially in time on one and only map. Empirical data collection is organised as outlined in Figure 2.3. Here, the overall model performance is assessed by comparing the model-generated scanpaths $\{\widetilde{\mathbf{r}}_F(1), \widetilde{\mathbf{r}}_F(2), \cdots\}$, with the actual ones $\{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}$.

The overall model dynamics can be described as follows. Given the i-th image stimulus at onset time t_0:

For all stages $k = 1, 2, 3$

Step 1 At time delay D_k, compute the model-based map $\mathcal{M}_k^{(i)}$

Step 2. Based on $\mathcal{M}_k^{(i)}$, generate "subject" fixations via the gaze shift model $\mathbf{r}_F^{(s,i)}(n) = f(\mathbf{r}_F^{(s,i)}(n-1), \mathcal{M}_k^{(i)})$:

$$\mathcal{M}_k^{(i)} \mapsto \{\widetilde{\mathbf{r}}_F^{(s,i)}(m_{k-1}+1), \cdots, \widetilde{\mathbf{r}}_F^{(s,i)}(m_k)\} = \mathcal{R}t_k^{(s,i)}, \tag{2.16}$$

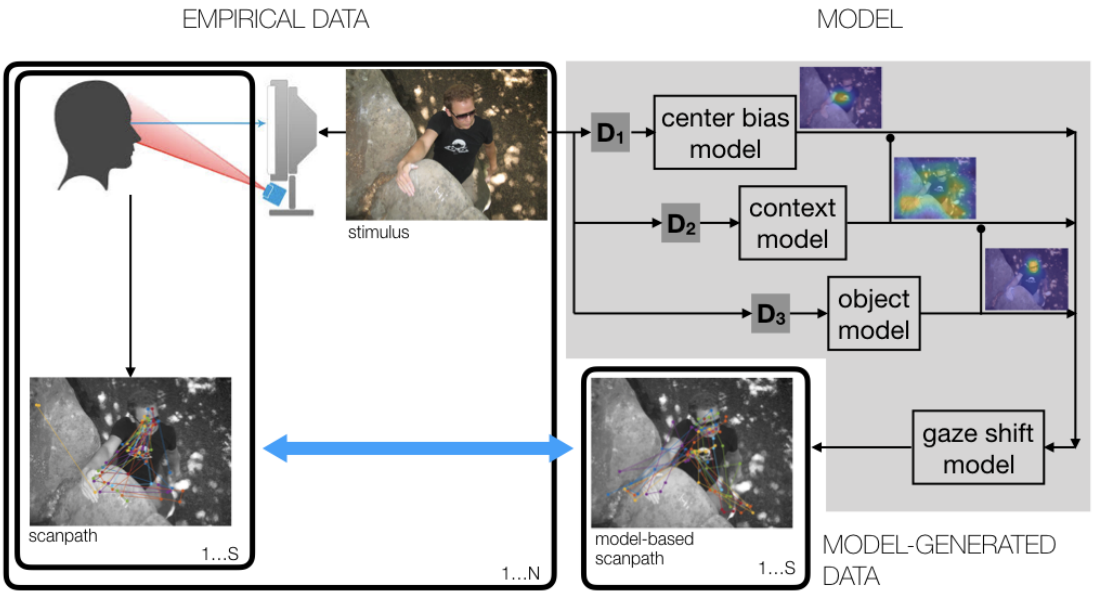


Figure 2.8: *The proposed three-stage model.*

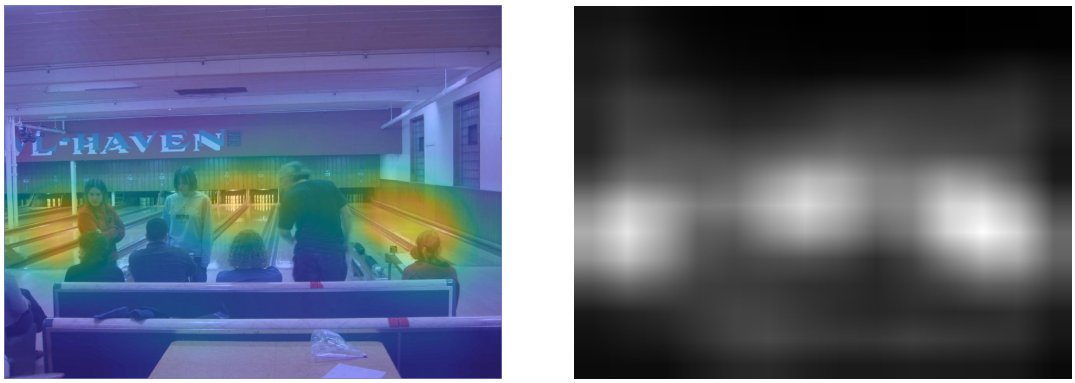
Eventually, collect the "time-aware" scanpath $\mathcal{R}t^{(s,i)} = \{\mathcal{R}t_1^{(s,i)}, \mathcal{R}t_2^{(s,i)}, \mathcal{R}t_3^{(s,i)}\}$.

For what concerns scanpath sampling, we again exploit the CLE (Boccignone and Ferraro, 2004) model.

More specifically we consider the following model components to compute the maps $\mathcal{M}_k^{(i)}, k = 1, 2, 3$:

1. **Center bias** Many studies (Tseng et al., 2009; Rothkegel et al., 2017) of attentional selection in natural scenes have observed that the density of the first fixation shows a pronounced initial center bias; this is modelled with a bidimensional Gaussian function located at the screen center with variance proportional to the image size, as shown in the first column of Fig. 2.6.

2. **Context model** Behavioural experiments (Oliva and Torralba, 2006) on scene understanding demonstrated that humans are able to correctly identify the semantic category of most real-world scenes even in case of fast and blurred presentations. Therefore, objects in a scene are not needed to be identified to understand the meaning of a complex scene. The rationale presented in Oliva and Torralba (2006), where a formal approach to the representation of scene *gist* understanding is presented, was further developed in Zhou et al. (2017) addressing scene classification via CNNs. The models were trained on the novel Places database consisting of 10 million scene photographs labelled with environment categories. In particular, we exploited the WideResNet (Zagoruyko and Komodakis, 2016) model fine-tuned on a subset of the database consisting of 365 different scene categories. The context map, therefore, is the result of the top-1 predicted category Class Activation Map (CAM) (Zhou et al., 2016). CAM indicates the discriminative image regions used by the network to identify a particular category and, in this work, simulates the exploration phase during which observers look at those



(a) *Scene, predicted as "bowling_alley"* **(b)** *Context map*

Figure 2.9: *Components of the context map. In (a) is shown the Class Activation Map of a scene correctly identified as "bowling alley", while in (b) the corresponding considered context map*

portions of the image which are supposed to convey the relevant information for the scene context understanding.

In Figure 2.9 is shown an example extracted from the dataset described in Judd et al. (2009), where a bowling alley is correctly identified by the network when focusing on the bowling lanes.

3. **Object model** The last stage to the realization of the final scanpath accounts for the convergence of fixations on relevant objects.

   It is worth noting that the relevance of an object is in principle strictly related to a given task (Tatler et al., 2011). The study presented here relies on eye-tracking data collected from subjects along a free-viewing (no external task) experiment and the sub-model design reflects such scenario. However, even under free-viewing conditions, it has been shown that at least faces and text significantly capture the attention of an observer (Cerf et al., 2009). Clearly, when these kinds of object are missing, other common objects that might be present within the scene become relevant.

   In order to obtain a realistic object map we exploited three different sub-frameworks implementing face detection, text detection and generic object segmentation, respectively. The output of each detector contributes, with different weight, to the final object map.

   More specifically, the face detection module relies on the HR-ResNet101 network (Hu and Ramanan, 2017) that achieves state-of-the-art performance even in presence of very small faces. This extracts canonical bounding box shapes that identify the regions containing a face. An example of the face detection phase is provided in Figure 2.10a.

   The generic object detection component is implemented via Mask R-CNN (He et al., 2017; Girshick et al., 2018). The latter capture objects in an image, while simultaneously generating a high-quality segmentation mask for each instance.

(a) *Faces (0.5)* **(b)** *Objects (0.1)* **(c)** *Text (0.4)*

(d) *Object map*

Figure 2.10: *Components of the object map: (a) shows the result of the face detector module; (b) the result of the object segmentation; (c) text detection result. In brackets, the weights of each component, in terms of contribution to the final object map (d).*

The CNN is trained on the COCO dataset (Lin et al., 2014), that consists of natural images that reflect everyday scene and provides contextual information. Multiple objects in the same image are annotated with different labels, among a set of 80 possible object categories, and segmented properly. Figure 2.10b shows an example, where all persons present in the image, as well as traffic lights and cars are precisely identified and segmented. The text detection component is represented by a novel Progressive Scale Expansion Network (PSENet) (Li et al., 2018), which can spot text with arbitrary shapes even in presence of closely adjacent text instances. An example of text detection result is shown in Figure 2.10c.

Simulation The evaluation procedure is eventually the following: we generated four different maps for each image $\mathbf{I}^i$ of the dataset. Three of these are the results of the adopted sub-models: center bias, context and object. The latter is obtained by combining the outputs of the three detectors: faces, text and common objects. The first two are the most relevant cues (Cerf et al., 2009) and we empirically assigned weights 0.5 and 0.4, respectively, while weighting 0.1 the object segmentation result. The final object map is later normalized. Such weighting allows to attribute more importance (saliency) to a specific object w.r.t. the others. As a consequence, the objects with higher weights will have higher probability of being chosen as candidates for a simulated fixation. The

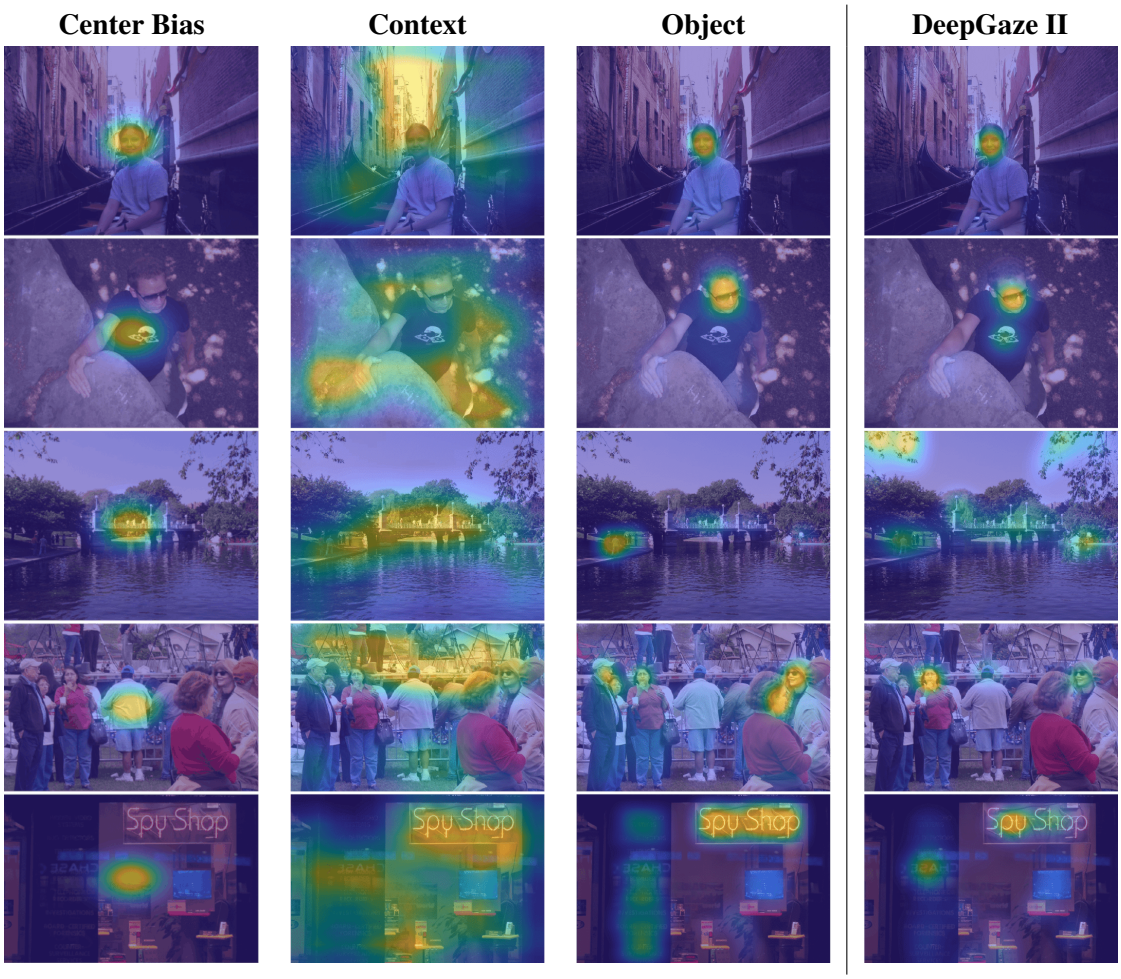


Figure 2.11: *Example of different maps generated for five images extracted from MIT1003 dataset. From left to right: the center bias, the context map and the object map, superimposed on the original stimulus; the saliency map resulting from saliency model DeepGaze II.*

adopted weights are those that yielded the best results empirically.

The comparison was carried out with the state-of-the-art static saliency model DeepGaze II (Kummerer et al., 2017).

All the considered saliency maps are convolved with a Gaussian kernel with $\sigma = 35$px (corresponding to 1dva for the MIT1003 dataset). Fig. 2.11 shows examples of the generated maps.

These were used to support the generation of $N_S = 15$ scanpaths for both the proposed and DeepGaze II approach, via the CLE gaze shift model (Boccignone and Ferraro, 2004). The number of fixations generated for each subject is sampled from the empirical distribution of the number of fixations performed by the human observer over each stimulus. Furthermore, in the proposed model, the switching time from the center bias map to the context map is set to 500 ms, while the permanence of the second map is equal to 1000 ms and the sampling of fixations from the object map is done for 1500 ms. In terms of delay time D_m, each model m is activated at $D_m = \{0, 500, 1000\}$ms,

	ScanMatch	Determinism	CORM
DeepGazeII w/o CB	0.34 (0.10)	41.16 (16.23)	19.09 (6.21)
DeepGazeII w/ CB	**0.41** (0.07)	50.34 (13.04)	16.39 (4.22)
Ours	0.36 (0.06)	**54.47** (6.54)	**13.75** (2.65)
Ground truth	0.45 (0.05)	59.72 (7.64)	10.02 (2.11)

Table 2.2: *Average values (standard deviations) of the considered metrics evaluated over all the artificial and human "observers" related to the same images in the dataset.*

while inhibiting the output of model $m - 1$.

Figure 2.12 shows CLE generated scanpaths, compared against the actual set of human scanpaths. The examples show how considering the context in the exploration of a scene and the precise detection of salient high-level objects, leads to scanpaths that are closer to those resulting from human gaze behaviour, than scanpaths generated via the classic saliency map. In particular, the first two rows of Fig.2.12 show how the contribution of the context map reflects the human exploration of the background, rather than focusing only on faces. The third row shows an example where DeepGaze II gives high relevance to low-level features that are not salient for human observers. In the following row it can be noticed how during the exploration phase all the faces are relevant, even when these are not faced towards the observer. Finally, as regards text, the last example shows how the whole text region is relevant and not just individual portions of it.

To quantitatively support such insights, the generated scanpaths have been evaluated on each image of the dataset by adopting the same metrics used in the previous section, namely ScanMatch and RQA.

Results All the generated scanpaths belonging to our approach and DeepGaze II have been evaluated against human scanpaths for each image. Table 2.2 reports the average values over all the "observers" related to the same images in the dataset. To quantify the intra-human similarity, an additional measure resulting from the comparison of ground truth scanpaths with themselves is provided.

It must be noted that, in case of DeepGaze II, the adopted model is fine-tuned exactly on the same dataset adopted for testing. Although this clearly introduces bias on the results, it can be seen how the proposed approach outperforms the model without center bias in all three considered metrics. When comparing with the "center bias-aware" model, the ScanMatch result of our approach is worse. In this case, the DeepGaze II output benefits from the addition of a prior distribution estimated over all fixations from the test dataset. On the other hand, our model has been created using a principle driven approach that does not need to use this same dataset.

2.5 A glimpse through the lenses of probability

The previous sections put the accent on the fact that a proper model of eye guidance must be able to account for the moment-to-moment relocation of gaze; in a nutshell it should aim at answering the question *Where to look next?*. As shown earlier, answering

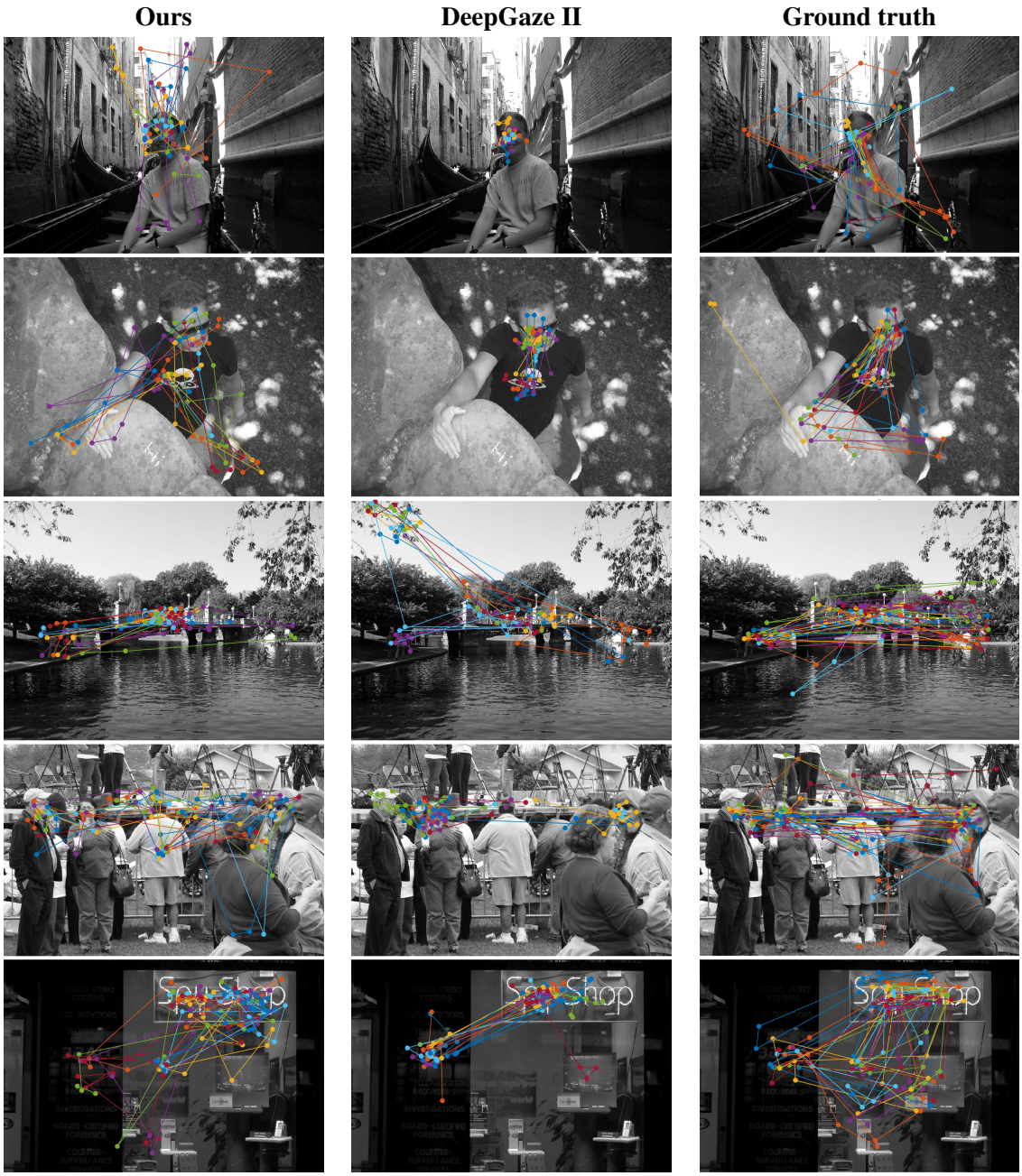


Figure 2.12: *Examples of scanpaths for the images considered in Fig. 2.11. Left to right: 15 model-generated scanpaths, from the proposed method, 15 model-generated scanpaths from the DeepGaze II saliency map, 15 scanpaths from actual human fixation sequences (ground-truth). Different colours encode different "observers", either artificial or human.*

this question hides many hitches. The dynamic nature of gaze together with its random components and the systematic tendencies make modelling gaze shifts a challenging problem.

For all this reasons, a convenient way to phrase a model of attentive eye guidance is in the language of probabilities. In this vein, Tatler and Vincent (2009) proposed to re-define the process of gaze relocation in terms of the posterior probability density function $P(\mathbf{r}|\mathcal{W})$, representing the probability of performing the gaze shift $\mathbf{r} = \mathbf{r}_F(t) - \mathbf{r}_F(t-1)$ given the perceptual representation $\mathcal{W}$, where $\mathbf{r}_F(t)$ represents the gaze position at time t. By means of the *Bayes rule*, this quantity can be decomposed into simpler components:

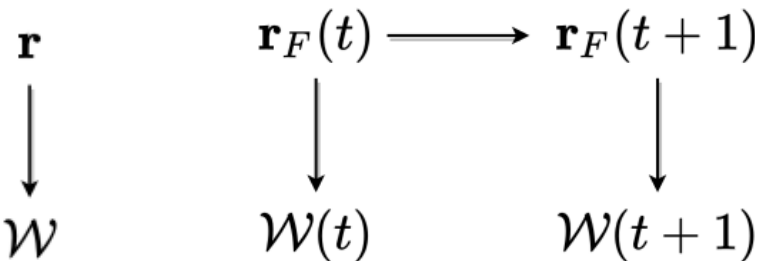
$$P(\mathbf{r}|\mathcal{W}) = \frac{P(\mathcal{W}|\mathbf{r})}{P(\mathcal{W})} P(\mathbf{r}) \tag{2.17}$$

Bayes rule, allows to break down the posterior into a *likelihood* $P(\mathcal{W}|\mathbf{r})$, a *prior* distribution $P(\mathbf{r})$ and the *marginal likelihood* or *evidence* $P(\mathcal{W}) = \int_{\mathbf{r}} P(\mathcal{W}|\mathbf{r})P(\mathbf{r})$. Crucially, the likelihood describes how the perceptual representation $\mathcal{W}$ might be involved in choice of the gaze shift $\mathbf{r}$. This quantity is normalized by $P(\mathcal{W})$, the evidence of the perceptual representation. Intuitively, this rapport controls for the natural abundance of a particular cue of the perceptual representation; for instance, as Tatler and Vincent (2009) put it: "*if yellow items are commonly fixated then one may initially infer that yellow items predict fixations, but if yellow items are very common in the scene then yellow is a less effective predictor of eliciting fixations*".

Remarkably, this approach is a very general one; indeed $P(\mathcal{W})$ can come from a variety of data sources such as simple feature cues, derivations such as Itti's definition of salience, object-or other high-level sources.

The second term on the r.h.s. of Equation 2.17 is the pdf $P(\mathbf{r})$ incorporating prior knowledge on gaze shift execution. In other words, this quantity accounts for all such properties of the gaze shifts that are independent from the stimuli, like the systematic tendencies and biases described earlier.

Figure 2.13a shows the generative model behind Equation 2.17 in the form of a simple Probabilistic Graphical Model (PGM), while Figure 2.13b depicts the same PGM unrolled in time. Note that now the arc $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t+1)$ makes explicit the dynamical nature of gaze shifts.



(a) *PGM describing Equation 2.17* **(b)** *PGM describing Equation 2.17 unrolled in time*

Figure 2.13: *The generative models behind Equation 2.17*

According to the model depicted in Figure 2.13, if all the probability distribution

are given, the attentive process can be simulated through *ancestral sampling*, namely:

1. Sample the gaze shift from the prior: $\mathbf{r}^* \sim P(\mathbf{r})$
2. Sample the perceptual representation of the environment, given the gaze shift: $\mathcal{W}^* \sim P(\mathcal{W} \mid \mathbf{r}^*)$

Conversely, in the vein of a typical computational model, sampling a new gaze shift given the perceptual representation is achieved by solving the Bayes' rule (Equation 2.17).

2.5.1 The unifying view of the attentive process

As a matter of fact Equation 2.17 summarizes all the features of the prototypical computational model of attentive eye guidance:

1. it handles the variability in a principled way by means of probabilities
2. it provides a rigorous definition of the systematic tendencies and biases of eye movements through the bayesian prior
3. it defines gaze shifts as a dynamical process

As pointed out in previous sections this last issue has been often overlooked. As a consequence, many models are more effectively spelled in the probabilistic framework as follows:

$$P(\mathbf{r}_F \mid \mathcal{W}) = \frac{P(\mathcal{W} \mid \mathbf{r}_F)}{P(\mathcal{W})} P(\mathbf{r}_F) \tag{2.18}$$

Now the posterior $P(\mathbf{r}_F \mid \mathcal{W})$ represents the probability of gazing at position $\mathbf{r}_F$ rather then the probability of performing the gaze shift $\mathbf{r} = \mathbf{r}_F(t) - \mathbf{r}_F(t-1)$, hence dynamics is not taken into account. In other words, two gaze position that follow one another over time are assumed to be independent:

$$P(\mathbf{r}) = P(\mathbf{r}_F(t) - \mathbf{r}_F(t-1)) \simeq P(\mathbf{r}_F(t) \mid \mathbf{r}_F(t-1)) = P(\mathbf{r}_F(t)) \tag{2.19}$$

It is interesting to note that such definition encapsulate the vast majority of approaches relying on statistical machine learning and more in general all such techniques belonging to the realm of saliency prediction as a sub-field of computer vision. In particular by substituting $\mathbf{r}_F$ with a random variable $\mathbf{L}$ denoting a location in the scene and $\mathcal{W}$ with $\mathbf{F}$ denoting features (of any kind), then Equation 2.18 boils down to:

$$P(\mathbf{L} \mid \mathbf{F}) = \frac{P(\mathbf{F} \mid \mathbf{L})}{P(\mathbf{F})} P(\mathbf{L}) \tag{2.20}$$

If we consider $\mathbf{L}$ as a binary random variable taking values in $[0, 1]$, then $P(\mathbf{L} = 1 \mid \mathbf{F})$ represents the probability for a particular location of the stimuli (either a pixel, a super-pixel or a broader region) to be classified as salient given that the feature $\mathbf{F}$ has been observed. If no prior knowledge about the saliency of the stimuli is assumed (i.e.

$P(\mathbf{L})$ is the uniform distribution), then $P(\mathbf{L} = 1 \mid \mathbf{F}) \simeq P(\mathbf{F} \mid \mathbf{L} = 1)$ so that saliency of $\mathbf{L}$ can be predicted by maximization of the likelihood function $P(\mathbf{F} \mid \mathbf{L} = 1)$ which, for instance, can be estimated through Kernel Density Estimation (Seo and Milanfar, 2009). More generally the ratio $f(\mathbf{L}) = \frac{P(\mathbf{L}=1|\mathbf{F})}{P(\mathbf{L}=0|\mathbf{F})}$ can be considered, thus casting the problem of saliency prediction as a classification one. A plenty of work proposed in the literature follow this rationale (Kienzle et al., 2006; Judd et al., 2009; Yan et al., 2010; Lang et al., 2012; Jiang et al., 2015a; Harel et al., 2007; Yu et al., 2014; Mathe and Sminchisescu, 2015; Vig et al., 2014); this includes even more modern techniques based on deep learning (Borji, 2019), which as a matter of fact do not bring any conceptual novelty as to the use of Equation 2.20.

Interestingly enough, the model described in equation 2.20 can be further simplified by setting the likelihood and prior terms to constants ($P(\mathbf{F} \mid \mathbf{L}) = const$ and $P(\mathbf{L}) = const$), thus obtaining:

$$P(\mathbf{L} \mid \mathbf{F}) = \frac{1}{P(\mathbf{F})} \tag{2.21}$$

This new minimal model (Eq. 2.21) now states that the probability of fixating a location $\mathbf{L}$, having observed features $\mathbf{F}$ is higher the more unlikely the feature $\mathbf{F}$ is (unlikeliness defined as $\frac{1}{P(\mathbf{F})}$). To come back to the pictorial example given by Tatler and Vincent (2009), if yellow items are very uncommon in the scene, once they are detected suddenly capture attention, thus may be considered as good predictors of fixation location. This is exactly what bottom-up saliency based models assume by detecting high contrast regions (with respect to either luminance, color, texture or motion). In other words the popular model by Itti et al. (1998) is well described probabilistically by Eq. 2.21.

Such probabilistic view of computational models allows to rephrase more rigorously the criticism about the lack of dynamics in many models of attention. This is particularly true for the most exploited aspect of this models: saliency prediction. As a matter of fact, surmising the absence of dynamics in the attentive process is a strong assumption that clearly overlooks the broader aspects of attention deployment that are, instead, provided by the unfolding in time of eye movements. To sum up, the gaze deployment process can be defined as a dynamical system, that is subject to stochasticity; such definition clearly matches with that of a *stochastic process*. Under such rationale, in the next chapter we provide a brief introduction to stochastic processes and their application to modelling eye movements.

2.6 Summary

In this Chapter we give a general overview of the computational modelling of attentive eye guidance. We start from a brief description of the early approaches, so to continue with the more modern techniques mainly relying of machine learning and deep learning models. We advance a criticism to such late approaches grounded on the lack of modelling of the dynamics of the attentive process. Consequently, a novel model of time-aware scanpath generation on static stimuli is proposed. The Chapter ends with a probabilistic description of visual attention models.

CHAPTER 3

Stochastic Processes, Eye Movements and Ecology

WHEN it comes to the description of natural phenomena, one unavoidably has to face the problem of randomness at some point. Indeed, many of the processes pertaining nature are fluctuating ones, thus can be marked as carriers of some degree of uncertainty.

For instance, one could think about some easily measurable quantities like the hourly series for the temperature of a city during a week, the progress of stock markets or the amplitude of a sound signal. These are all examples of fluctuating phenomena for which we are able to build predictive models that forecast future values, but with a certain amount of error. We may ask, to what extent we are able to reduce such error. Is it possible to obtain perfect future forecasts? Answering such question calls into play the philosophical debate about the nature of these fluctuations; are they intrinsic properties of the observed phenomena, or are our way, as scientists, to take into account what we are not able to measure. In other words, is uncertainty part the generating process, or it's just an admission of ignorance?

According to the beliefs of science up to the nineteenth century, there is no particular reason to talk about randomness associated to the quantity itself; rather, the value of any given variable of interest is assigned by nature according to an extremely complex but deterministic procedure. It this respect, we should be able to forecast the future values of such variables with certainty, if we have access to all the relevant information. In this deterministic picture of nature, what we consider *noise*, randomness or uncertainty evolves from our ignorance of the boundary conditions. This is the gist of the *scientific determinism* as interpreted by Pierre-Simon Laplace, who states the existence of an "intellect" (later referred to as the Laplace's daemon) that *"at a certain moment would*

know all forces that set nature in motion" (marquis de Laplace, 1902).

This conception was later subverted by quantum theory on the one hand and chaos theory on the other (a deterministic system can exhibit an unpredictable behavior as in the butterfly effect) and we now know that the purely statistical element is an essential basis of our world (Gardiner, 2011). The very same concept of *Laplace's daemon* was recently disproved by means of Turing machines under the assumption of free will (Rukavicka, 2014).

The non determinism of nature is not a surprising thing; in fact the stochastic description of natural phenomena more easily meets our every day experience in which quantities are predictable to some extent but not completely. Hence the modelling of such fluctuating phenomena calls for a statistical explanation or, in other words, stochastic models.

The beginning of stochastic modelling of natural phenomena, can be traced back to the Einstein's explanation of the theory of Brownian Motion (Einstein, 1905). Brownian Motion is the name that was given to the animated and irregular state of motion exhibited by small pollen grains suspended in water after the pioneering work of the botanist Robert Brown in 1827. Brown was the first who performed a systematic investigation of the phenomena, but the first mathematical description of it was given almost eighty years later by Einstein (1905) and Von Smoluchowski (1906).

As a byproduct, Einstein's work laid down the basis for a bunch of "tools" that were lately further developed in a more general and rigorous way and that nowadays are considered the fundamental concepts of the theory of stochastic processes (many of which will be treated in the present Chapter). Things like the Markov assumption, the *Chapman-Kolmogorov Equation* (the central dynamical equation of all Markov processes), or the diffusion equation describing the behaviour of an ensemble of particles (later evolved in the *Fokker-Planck Equation*), are all reverberations of Einstein's dissertation.

Einstein's seminal paper was of inspiration to a later work of Paul Langevin which came up with a new way of deriving the same results presented in the Einstein's work. This led to the development of the renowned Langevin equation describing the dynamics of a single particle, providing the first example of a *Stochastic Differential Equation* (SDE). Langevin's work was later improved and expanded by Ornstein and Uhlenbeck (Uhlenbeck and Ornstein, 1930) who came up with the namesake process describing the velocity of a Brownian particle.

Besides the description of the physical phenomena for which they were conceived, stochastic processes turned out to be useful in a variety of different fields, from the analysis of electrical circuits and radio wave propagation to the modelling of stock markets in finance, (e.g. the Black-Scholes (Black and Scholes, 1973) or Vasicek (Vasicek, 1977) models). Moreover, stochastic processes were successfully used to model biological systems, indeed the first account of Brown's work was to find out if the movements of the pollen particle where a manifestation of life (an hypothesis that was promptly ruled out by experiments). Stochastic models of gene expression, models of fluctuations in bacteria's protein concentration and crucially models of animal movements are all realizations of random phenomena.

Of key interest for the present thesis is the concept of super-diffusive processes, which were firstly brought to the attention of the scientific community by the work of

Richardson (1926), which presented data exhibiting a behaviour in contradiction with normal diffusion that could be explained by a deviation of the statistic of fluctuation from the Gaussian distribution. Such "anomalous diffusion" was then rediscovered in many other fields like finance, biology and ecology mainly under the mathematical description of Lèvy Flights. Interestingly enough the latter were employed by Brockmann and Geisel (2000) to describe the oculomotor behaviour of humans on images, while establishing at the same time, a connection between eye movements and Ecology (cfr Section 3.5.2).

Crucially, eye movements recorded from human subjects while watching a stimuli, can be conceived as different manifestations of the same fluctuating phenomena, namely the result of visual attention allocation on the scene. Hence, they lend themselves to be described via stochastic models. In particular, each sequence of gaze positions can be associated to a different realization of a stochastic process.

The forthcoming sections of this Chapter, aim at expanding the concepts shortly provided by this introduction, starting from the rigorous definition of stochastic processes and their description, their adoption as sound models of eye movements, up to the introduction of the foraging perspective.

3.1 Stochastic Processes

A stochastic process is a collection of random variables $\mathbf{X}(t)$ indexed by the variable t, usually denoting time. If t takes values in the set of real numbers, then $\mathbf{X}(t)$ is a *continuous time stochastic process*; on the other hand, if t belongs to natural numbers then $\mathbf{X}(t)$ is a *discrete time stochastic process*.

Figure 3.1, depicts different eye movements as recorded from a group of observers while looking at the same image. We can treat each trajectory as a realization of a stochastic process $\mathbf{X}(t)$. Observing a realization, means attributing a specific value to each random variable:

$$\mathbf{X}(t_1) = \mathbf{x}_1, \quad \mathbf{X}(t_2) = \mathbf{x}_2, \quad \mathbf{X}(t_3) = \mathbf{x}_3, \cdots, \tag{3.1}$$

In this specific case, $\mathbf{X}(t)$ is a two dimensional random variable representing the i and j coordinates of the position of the eye, hence at time t_k, $\mathbf{X}(t_k) = \mathbf{x}_k = [i, j]$.

We can observe the process $\mathbf{X}(t)$ in time by considering a single realization (one specific scanpath), or we could consider the ensemble of trajectories and look at the empirical distribution of the variable $\mathbf{X}(t_n)$ at time t_n; this can be conceived as an approximation of the true PDF $P(\mathbf{x}_n, t_n)$, answering the question *What's the probability of looking at a specific region of the stimuli at time t_n?*.

This example highlights the fundamental difference between a stochastic variable and a stochastic process; if the former describes the probability of having looked at a specific region of the stimuli at fixed time t_n, the latter describes the whole sequence of eye positions as a set of random variables $\mathbf{X}(t_1), \mathbf{X}(t_2), \cdots, \mathbf{X}(t_n), \cdots$ put in appropriate order.

Put simpler, one could consider a one-dimensional stochastic process $X(t)$, representing, for instance, the i coordinate of a scanpath. This is depicted in Figure 3.2. Each horizontal slice may be conceived as a realization of the stochastic process underling the i coordinate of a scanpath. By fixing the time variable $t = t_n$ we are intuitively

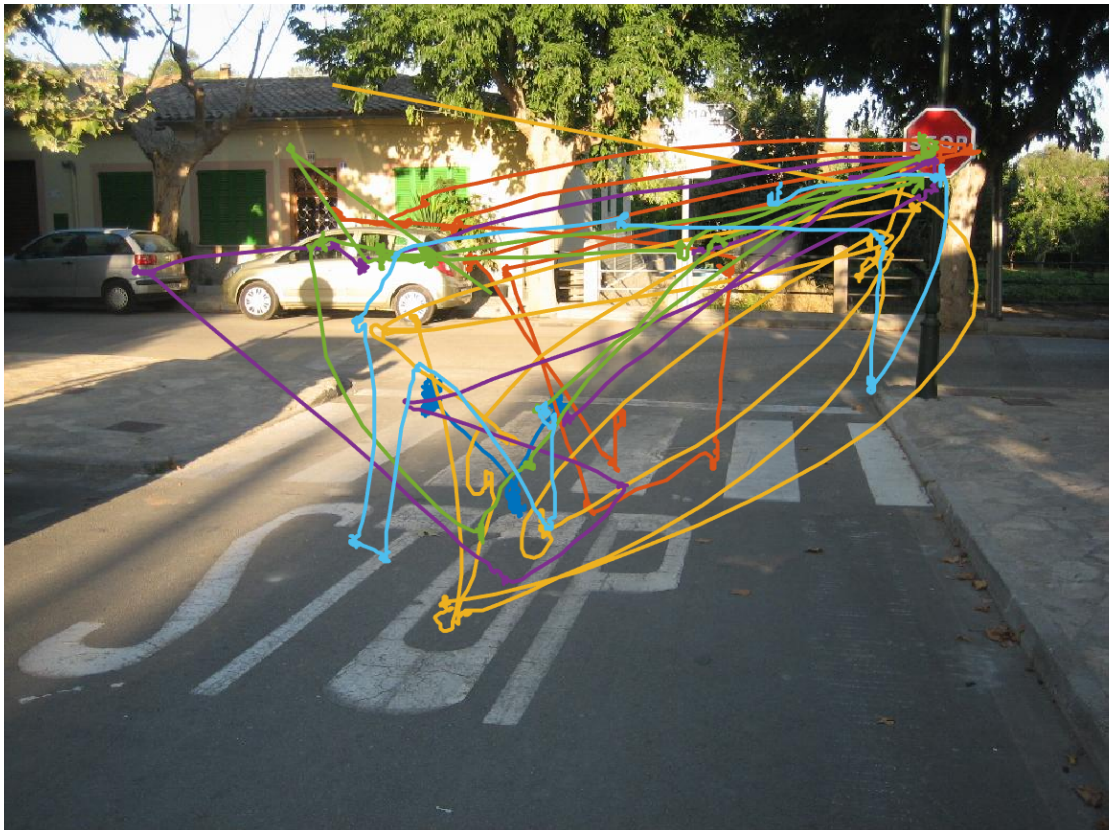


Figure 3.1: *Eye movements recorded from a set of different observers on the same stimuli from the dataset Judd et al. (2009). Each color corresponds to a different observer*

slicing vertically the ensemble of realizations, thus obtaining an empirical estimate of $P(x_n, t_n)$.

Mathematically, a stochastic process is completely defined by its joint probability density function; if $P(x_1, t_1)$ defines the PDF for $X(t_1)$ at time t_1, $P(x_1, t_1; x_2, t_2)$ is the joint probability density for variables $X(t_1)$ and $X(t_2)$ of the stochastic process at times t_1 and t_2, respectively. In other words, this quantity represents the probability that the random variable $X(t_1)$ at time t_1 takes a value x_1 *and* the random variable $X(t_2)$ at t_2 takes the value x_2. Similarly, the k-th joint PDF at successive times t_i (with $i = 1, \cdots, k$) is $P(x_1, t_1; x_2, t_2; \cdots; x_k, t_k)$. The latter quantity completely specifies the statistical properties of the process.

The evolution over time of a stochastic process can be described via *transition probabilities*; these are nothing but the conditional probabilities of the future values of the process given the past. Such quantities can be easily obtained from the joint distribution via the product rule of probability, which allows to rewrite the joint as follows:

$$P\left(x_1, t_1; \cdots; x_k, t_k; \cdots; x_{k+l}, t_{k+l}\right) = P\left(x_1, t_1; \cdots; x_k, t_k\right) \times \\ P\left(x_{k+1}, t_{k+1}; \cdots; x_{k+l}, t_{k+l} \mid x_1, t_1; \cdots x_k, t_k\right) \tag{3.2}$$

Hence, the transition probability writes:

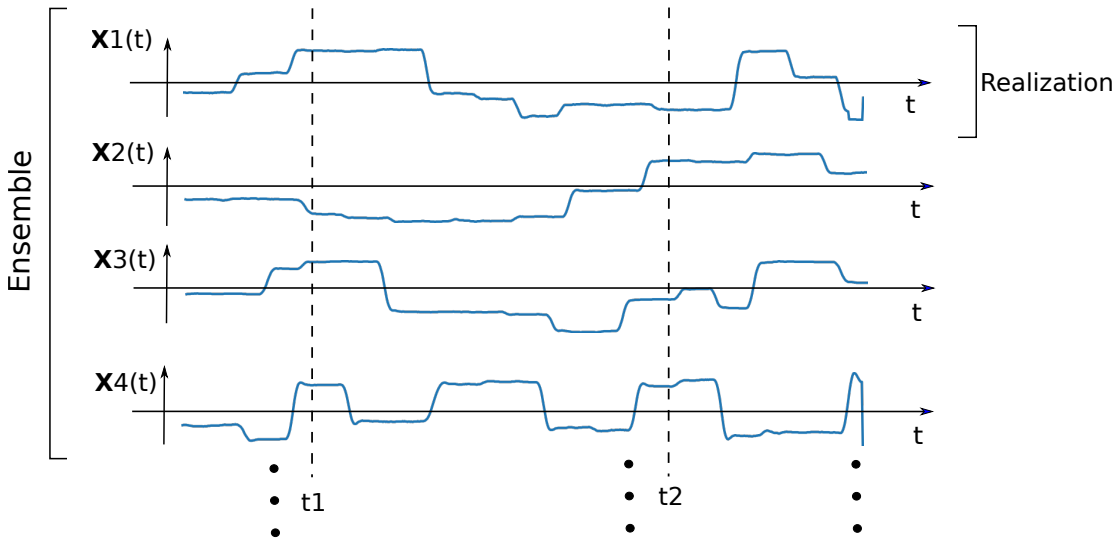


Figure 3.2: *Different realizations of a stochastic process*

$$P(\overbrace{x_{k+1}, t_{k+1}; \cdots ; x_n, t_n}^{\text{future}} \mid \underbrace{x_1, t_1; \cdots ; x_k, t_k}_{\text{past}}) = \frac{P(x_1, t_1; \cdots ; x_n, t_n)}{P(x_1, t_1; \cdots ; x_k, t_k)}. \tag{3.3}$$

The k-joint distributions can be reduced to the $(k-1)$-joint PDFs by integrating out the k-th variable:

$$\int_{\Omega_k} P(x_1, t_1; x_2, t_2; \ldots ; x_k, t_k)\, \mathrm{d}x_k = P(x_1, t_1; x_2, t_2; \ldots ; x_{k-1}, t_{k-1}) \tag{3.4}$$

where Ω_k is the support of x_k. By combining Equation 3.2 and 3.4, one obtains:

$$\begin{aligned} P(x_{k+1}, t_{k+1}; \ldots ; x_{k+l}, t_{k+l}) = \int_{\Omega_1} \ldots \int_{\Omega_k} P(x_1, t_1; \ldots ; x_k, t_k) \\ \times P(x_{k+1}, t_{k+1}; \ldots ; x_{k+l}, t_{k+l} \mid x_1, t_1; \ldots x_k, t_k)\, \mathrm{d}x_1 \ldots \mathrm{d}x_k \end{aligned} \tag{3.5}$$

By way of example one could consider the case of the two joint PDF in which Equation 3.5 becomes:

$$P(x_2, t_2) = \int_{\Omega_1} P(x_1, t_1)\, P(x_2, t_2 \mid x_1, t_1)\, \mathrm{d}x_1 \tag{3.6}$$

Here $P(x_2, t_2 \mid x_1, t_1)$ acts like the *evolution kernel* or *propagator* form (x_1, t_1) to (x_2, t_2). In other words, Equation 3.5 provides the explicit dependence of the joint on time. However, some kind of stochastic processes do not exhibit a time dependence, but eventually reach a situation in which their statistical properties remain unchanged in time, thus converging to a stationary bahaviour. Such processes are called *strict sense stationary* (SSS) and satisfy the relation:

$$P\left(x_1, t_1; x_2, t_2; \ldots; x_n, t_n\right) = P\left(x_1, t_1 + \tau; x_2, t_2 + \tau; \ldots; x_n, t_n + \tau\right) \tag{3.7}$$

for all $\tau > 0$ and n. Hence, for a strict sense stationary stochastic process $P\left(x_i, t_i\right) = P\left(x_i\right)$.

3.1.1 Summarizing a stochastic process

Given a realization of a stochastic process $x(t)$, the measurements that are readily available at any time t are quantities like the mean of the variance. In general, by computing standard summary statistic at any time, one assumes a particular shape of the noise (Gaussian, Lèvy, Poisson etc.). However such quantities do not provide any insight about the dynamics of the process, i.e. about the ability of the values measured at the current time to influence the future ones. The amount of dependence or memory of the measured signal can be characterized by the **autocorrelation function**:

$$C_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_0^T x(t)x(t+\tau)\mathrm{d}t \tag{3.8}$$

This can be seen as a time-average of the product $x(t)x(t+\tau)$ over the interval $(0, T)$ for a fixed value τ and for T tending to infinity. $C_{xx}(\tau)$ measures the correlation between two points of the signal $x(t)$ separated by a time shift τ . Large values of the autocorrelation function indicates that subsequent values of the process are strongly dependent on the previous ones; hence, our capability to predict the future values from its previous history increases.

Practically, when a finite and discretized signal is measured, the integral can be replaced by a sum and the average is computed on the total amount of sample collected (N); in this case, the autocorrelation function is approximated by the **sample autorrelation**:

$$c_{xx}(\Delta) = \frac{1}{N} \sum_{n=0}^{N-|\Delta|-1} x(n)x(n+\Delta) \tag{3.9}$$

Another way to summarize a stochastic process is by analysis of its spectral components by means of the **Power Spectral Density** (PSD):

$$S(\omega) = \lim_{T\to\infty} \frac{1}{2\pi T} \left| \int_0^T \mathrm{e}^{-\mathrm{i}\omega t} x(t)\mathrm{d}t \right|^2 \tag{3.10}$$

This function measures the contribution of each frequency component to the observed time series; interestingly enough, such quantity exhibits a tight connection with the autocorrelation function.

In fact, if strict-sense stationarity is a desirable property, it's often not met in many real-life processes. Fortunately, it is possible to show a "weaker" form of stationarity than the one defined above. One of the most common forms of stationarity that is used in practice is *wide-sense* (or *weak-sense*) stationarity (WSS). A random process is WSS if its mean function and its autocorrelation function do not change by shifts in time; that is, for each t_1, t_2 and τ:

$$\langle X(t_1)\rangle = \langle X(t_2)\rangle \tag{3.11}$$
$$\langle X(t_1)\,X(t_2)\rangle = \langle X(t_1+\tau)\,X(t_2+\tau)\rangle \tag{3.12}$$

Where $\langle\cdot\rangle$ denotes the expected value operator. Note that the first condition states that the mean function is not a function of time t, while the second condition states that the correlation function is only a function of the difference $t_2 - t_1$ and not t_1 and t_2 individually.

For WSS processes according to the Wiener-Khinchin theorem exists a conjugal relationship between the autocorrelation function $C_{xx}(\tau)$ and the PSD $S(\omega)$:

$$\begin{aligned} S(\omega) &= \frac{1}{2\pi}\int_{-\infty}^{-\infty} \exp(-i\omega\tau)C_{xx}(\tau)d\tau \\ C_{xx}(\tau) &= \int_{-\infty}^{-\infty} \exp(i\omega\tau)S(\omega)d\omega \end{aligned} \tag{3.13}$$

This means that the autocorrelation function can be estimated from the observed signal by means of the inverse Fourier Transform of the spectral density which is tipically easier to determine from the analysis of data series via the Fast Fourier Transform (FFT).

3.1.2 Markov Processes

The simplest stochastic process that can one can conceive is a process in where each random variable $X(t)$ is completely independent from the others. For such *Purely Random Process* the joint distribution can be easily written as the product of all the random variables that occur at each time instant:

$$P(x_1,t_1;\ldots;x_n,t_n) = P(x_1,t_1)\,P(x_2,t_2)\cdots P(x_n,t_n) \tag{3.14}$$

or, more succinctly:

$$P(x_1,t_1;\ldots;x_n,t_n) = \prod_{i=1}^{n} P(x_i,t_i) \tag{3.15}$$

Equation 3.15 states that the process $X(t)$ is completely memoryless and present no correlations. A particular case occurs when besides independence with respect to time t, the probability distributions of the random variables at each time instant $P(x_i,t_i)$ are governed by the same probability law. Such a process is said to be defined by a set of *independent and identically distributed* (i.i.d.) random variables. When $P(x_i,t_i)$ is assumed to be the Normal distribution, the popular *white noise* is recovered.

The autocorrelation function for the purely random process is given by:

$$C_{xx}(\tau) = A\delta(\tau) \tag{3.16}$$

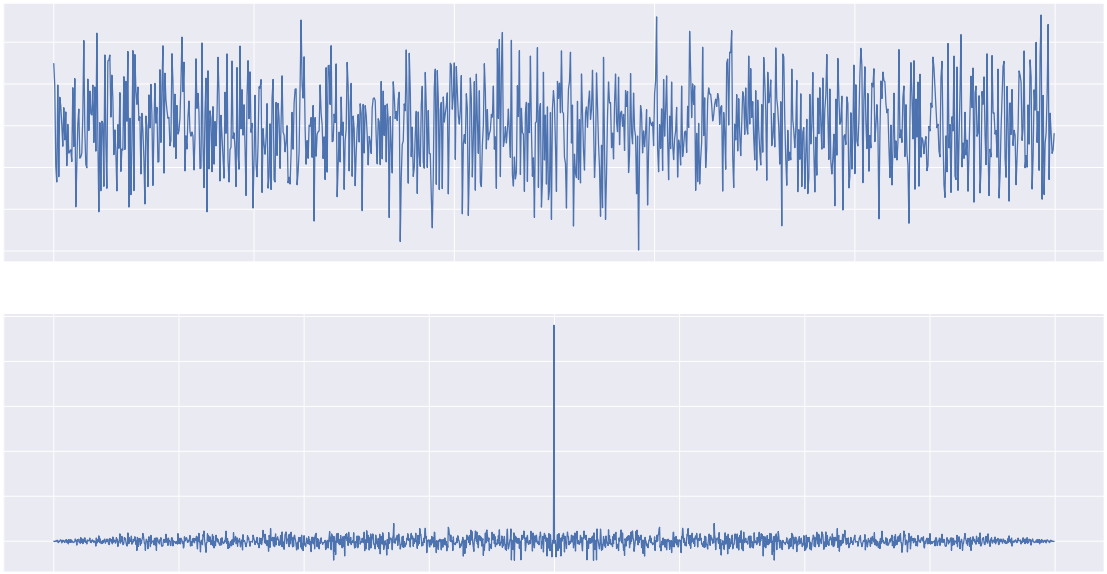


Figure 3.3: ***Top:*** *Purely random process (white noise).* ***Bottom:*** *Sample autocorrelation Function*

where $\delta(\cdot)$ is the Dirac Delta function. Eqaution 3.16 highlights the absence of correlations between the random variables that make up the process. This fact is empirically show in Figure 3.3, which depicts a simulation of a purely random process (white noise) with the corresponding sample autocorrelation function.

Unsurprisingly the sample autocorrelation function shows no dependence (correlation) between values as witnessed by its immediate decay.

More realistically, we know that the vast majority of processes in nature, despite being random, exhibit some degree of predictability or dependence between consecutive values. A step towards this direction is assuming that each value depends on the previous one only; this concept can be mathematically described by:

$$P\left(x_n, t_n \mid x_{n-1}, t_{n-1}; \ldots; x_1, t_1\right) = P\left(x_n, t_n \mid x_{n-1}, t_{n-1}\right) \tag{3.17}$$

with $t_1 < t_2 < \cdots < t_n$. Such hypothesis is called *Markovian Approximation* and such a process is called **Markov Process**. By assuming markovianity, the joint distribution can be extremely simplified by following Equation 3.2 and 3.17 as:

$$P\left(x_n, t_n; x_{n-1}, t_{n-1}; \cdots; x_1, t_1\right) = P\left(x_1, t_1\right) \prod_{i=2}^{n} P\left(x_i, t_i \mid x_{i-1}, t_{i-1}\right) \tag{3.18}$$

Hence, a Markov process is completely described by the initial distribution $P\left(x_1, t_1\right)$ and by the propagator $P\left(x_i, t_i \mid x_{i-1}, t_{i-1}\right)$.

An immediate example of a Markov process is the simple Random Walk, described by the following equation:

$$x_t = x_{t-1} + k\xi_t \tag{3.19}$$

here ξ_t is the noise term which is sampled from a suitable distribution (e.g. Gaussian, Bernoulli, etc.) and k is a scaling constant. If Equation 3.19 is iterated for a number of time steps (in this specific case, time is assumed to be discrete), one obtains a realization of the process (intuitively corresponding to one horizontal slice of figure 3.2).

ξ_t is a random variable which is sampled independently from the noise distribution $\xi_t \sim P(\xi)$. It is easy to see that the differences in sequential observations $x_t - x_{t-1} = \xi_t \sim P(\xi)$ are i.i.d. However, in the process described by Equation 3.19 the observations at each time step are not independent. Indeed the evolution in time of the variable at the current time step depends on the previous one.

For comparison with the purely random walk, Figure 3.4 shows a simulation of Equation 3.19 with the associated sample autocorrelation function. As notable, here the the process exhibits a much slower decay in the empirical autocorrelation function denoting the presence of memory.

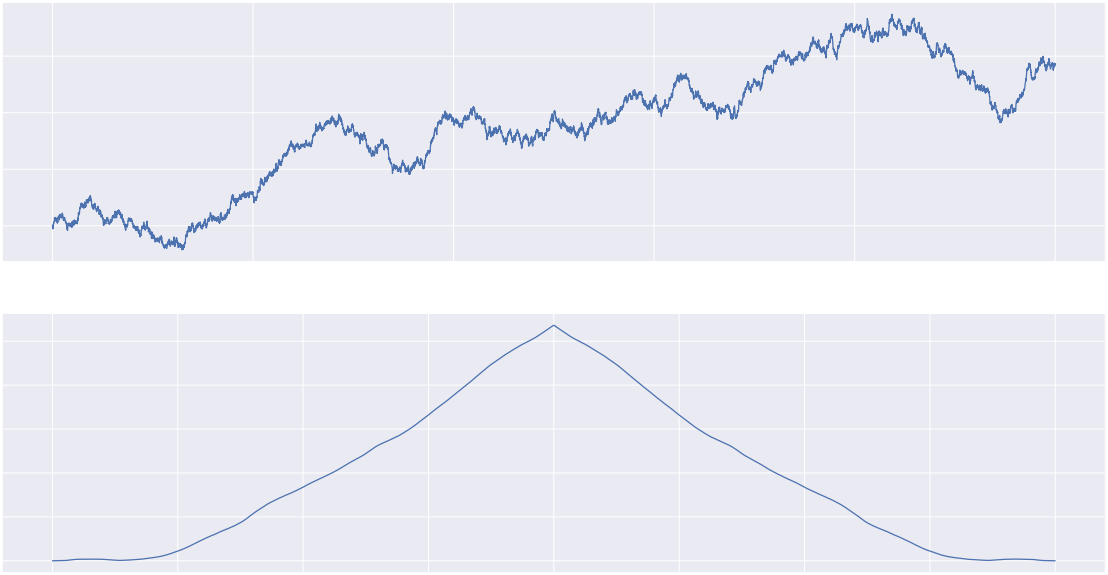


Figure 3.4: ***Top:*** *Simple Random Walk.* ***Bottom:*** *Sample autocorrelation Function*

If we assume the noise distribution $P(\xi)$ in Equation 3.19 to be Gaussian, that is $\xi \sim \mathcal{N}(0, \sigma^2)$, and we extend the process in two dimensions by defining:

$$\begin{aligned} x_t &= x_{t-1} + \xi_{x,t} \\ y_t &= y_{t-1} + \xi_{y,t} \end{aligned} \tag{3.20}$$

we obtain a simulation of the Brownian Motion (Figure 3.5) described in the introduction of the present chapter.

3.2 Levels of description of stochastic processes

Let's now head back to the depiction provided by Figure 3.1; this shows different scanpaths as recorded from different observers while attending the same image. As stated

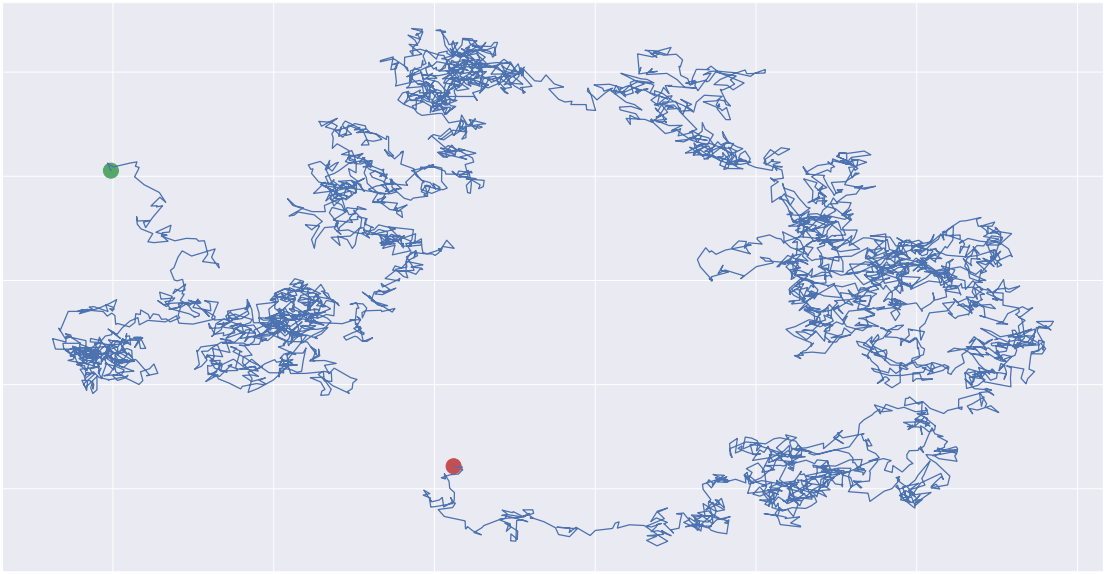


Figure 3.5: *Simulation of a simple random walk in two dimensions. Green ad red dots indicate the starting and end points of the simulation, respectively*

earlier, the main assumption here is to consider such trajectories as different realizations of the same stochastic process. The goal of building such a stochastic model is to be able to answer the question *"what's the probability* $P(\mathbf{x}, t)$ *of gazing at* $\mathbf{x}$ *at time* t*?"*.

This question can be addressed at different levels; for instance, one could be interested in considering the single scanpath as the tiniest unit of description. In this perspective, each observer can be conceived as a particle wandering in a two dimensional landscape (the image) like the brownian particle depicted in Figure 3.5.

By recording many observers, we are figuratively simulating many particles, each particle (scanpath) being governed by the same stochastic law. This is the **microscopic level** of description of a stochastic process in which the dynamics of a single particle is taken into account. Such view was proposed by the french physicist Paul Langevin, who gave rise to the mathematical field of the *Stochastic Differential Equations* (SDE) describing the behaviour of a single random walker.

Conceptually, many recordings from different observers may be conceived as many simulations of a SDE carried out independently, i.e. a Monte Carlo simulation. Hence, $P(\mathbf{x}, t)$ can be approximated by counting the number of times each position $\mathbf{x}$ has been gazed at a given time t.

Conversely, one could consider directly the evolution of the probability density $P(\mathbf{x}, t)$ of a many particles system (ideally an infinite number of particles), without actually taking care of the behaviour of the single random process, but considering the bigger picture, the collective phenomenon. This was the approach pursued by Einstein when derived the diffusion equation on brownian particles. This coarse-grained representation is the **macroscopic level** description and relies on the so called *Master Equation* and *Fokker-Plank Equation*.

The connection between the microscopic and the macroscopic levels is provided by the **mesoscopic level** of description. In the picture outlined thus far, this is represented

by the *Chapman-Kolmogorv Equation* (C-K Equation). This equation states that the probability of a particle to be in position $\mathbf{x}+\Delta\mathbf{x}$ at time $t+\Delta t$ is given by summing all possible displacements $\Delta\mathbf{x}$ multiplied by the probability of being at $\mathbf{x}$ at time t. Note how this relies on the assumption that each displacement is independent of the previous history; we are thus assuming markovianity.

Crucially, the C-K Equation is the mathematical tool that allows to coarse-grain the representation from the microscopic level (SDE) to the macroscopic one (Master Equation and Fokker-Plank Equation). This concepts are further developed in the following.

3.2.1 Microscopic Level

The microscopic level of description of a stochastic process involves the definition of the dynamical law of the single realization of the process. The typical tool employed for describing a dynamical system is the differential equation, which usually comes in the form:

$$\frac{dx(t)}{dt} = a(x(t), t) \tag{3.21}$$

Here, the variable x represents the quantity whose dynamics we wish to characterize. Intuitively Equation 3.21 describes the rate of change of x w.r.t. the variable t. In other words it provides the state of x at the next time step given the current state. Clearly, the law formalized in Equation 3.21 is deterministic; if noise or uncertainty has to be taken into account, then the evolution equation becomes a *Stochastic Differential Equation* (SDE) and writes:

$$\overbrace{\frac{dx(t)}{dt}}^{\text{state-space rate of change}} = \overbrace{a(x(t),t)}^{\text{deterministic comp.}} + \overbrace{b(x(t),t)\xi(t)}^{\text{stochastic comp.}} \tag{3.22}$$

This is called *Langevin Equation*, after the definition given by Paul Langevin to describe the motion of a brownian particle.

Notably, Equation 3.22 is composed by a deterministic component $a(x,t)$, usually called *drift*, and a stochastic one $b(x,t)\xi(t)$, called *diffusion*. Here $\xi(t)$ is the noise sampled from a probability density function $\xi(t) \sim P(\xi)$, usually a Normal distribution.

The Langevin equation, written as in Equation 3.22, poses some formal problems; indeed $\xi(t)$ is often obtained by sampling i.i.d. values from a PDF. As a consequence this may be non differentiable, thus making $x(t)$ non differentiable, too. This makes the left hand side of Equation 3.22 incoherent from such point of view. To overcome this problem, Equation 3.22 is usually written in the more mathematically sound form:

$$dx(t) = a(x(t),t)dt + b(x(t),t)\xi(t)dt = a(x(t),t)dt + b(x(t),t)dW(t) \tag{3.23}$$

where $W(t) = \int_0^t \xi(t')\,dt'$, so that the integration of the stochastic component $\int b(x,t)dW(t)$ can be performed according to the rules of stochastic calculus (in the Itô or Stratonovich sense (Higham, 2001)). In simpler terms, one can loosely interpret the quantity $dW(t)$ as an instance of a white noise process. An Itô SDE like the one described by Equation 3.23 can be written in its integral form, for all t and t_0, as:

$$x(t) = x(t_0) + \int_{t_0}^{t} a(x(t'), t')\, dt' + \int_{t_0}^{t} b(x(t'), t')\, dW(t') \tag{3.24}$$

Here $\int_{t_0}^{t} b(x(t'), t')\, dW(t')$ represents an Itô stochastic integral, stochastic generalization of the Riemannian integral in which the integrands and the integrators are stochastic processes.

One general way of simulating a SDE is to obtain a discretized version of it by taking a mesh of points t_i:

$$t_0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = t \tag{3.25}$$

and writing the SDE as:

$$x_{i+1} = x_i + a(x_i, t_i)\,\Delta t_i + b(x_i, t_i)\,\Delta W_i \tag{3.26}$$

where $x_i = x(t_i)$, $\Delta t_i = t_{i+1} - t_i$ and $\Delta W_i = W(t_{i+1}) - W(t_i) \propto \sqrt{\Delta t_i}\xi_i$.

The SDE can now be solved by recursively computing x_{i+1} given x_i and adding the drift and diffusion terms; notably such formulation highlights the markovian nature of the Langevin Equation. Such approximate procedure is called the **Euler-Maruyama discretization** and provides a general way of simulating a SDE.

3.2.2 Mesoscopic Level

Consider the 3-joint PDF $P(x_1, t_1; x_2, t_2; x_3, t_3)$. By means of Equation 3.3, this can be written as:

$$P(x_3, t_3; x_2, t_2 \mid x_1, t_1) = \frac{P(x_1, t_1; x_2, t_2; x_3, t_3)}{P(x_1, t_1)} \tag{3.27}$$

One of the most important properties of a Markov process, is obtained by marginalizing the conditional PDF $P(x_3, t_3, x_2, t_2 \mid x_1, t_1)$ with respect to x_2:

$$P(x_3, t_3 \mid x_1, t_1) = \int_{\Omega_2} P(x_3, t_3; x_2, t_2 \mid x_1, t_1)\mathrm{d}x_2 = \int_{\Omega_2} \frac{P(x_1, t_1; x_2, t_2; x_3, t_3)}{P(x_1, t_1)}\mathrm{d}x_2 \tag{3.28}$$

By exploiting the Markov property, Equation 3.28 writes:

$$P(x_3, t_3 \mid x_1, t_1) = \int_{\Omega_2} \frac{P(x_3, t_3 \mid x_2, t_2)\, P(x_2, t_2 \mid x_1, t_1)\, P(x_1, t_1)}{P(x_1, t_1)}\mathrm{d}x_2 \tag{3.29}$$

Hence:

$$P(x_3, t_3 \mid x_1, t_1) = \int_{\Omega_2} P(x_3, t_3 \mid x_2, t_2)\, P(x_2, t_2 \mid x_1, t_1)\, \mathrm{d}x_2 \tag{3.30}$$

Equation 3.30 is commonly known as the **Chapman-Kolmogorov (CK) Equation** and states that the transition probability from state $\{x_1, t_1\}$ to state $\{x_3, t_3\}$ occurs in two stages: one involving the transition from $\{x_1, t_1\}$ to $\{x_2, t_2\}$, and the second performing the move from $\{x_2, t_2\}$ to $\{x_3, t_3\}$. Intuitively, such process is carried out by

considering (integrating out) all the possible intermediate states $\{x_2, t_2\}$ that could act as transition states; this is depicted in Figure 3.6.

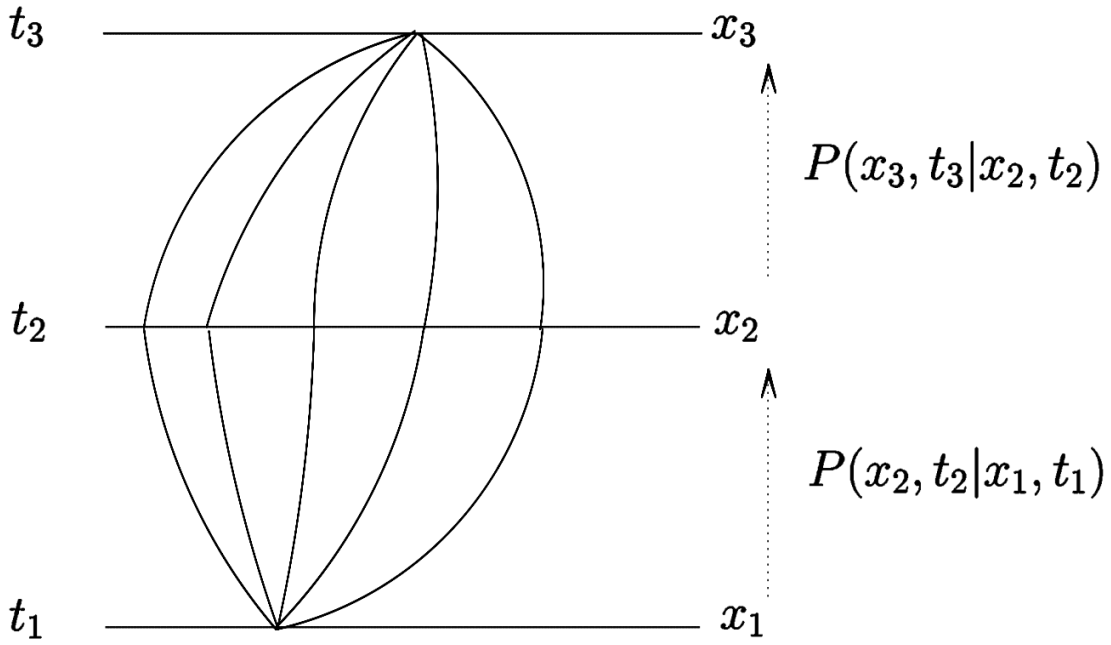


Figure 3.6: *Schematic picture for the Chapman-Kolmogorov equation (taken from Méndez et al. (2014))*

The CK Equation is a consistency equation for the conditional probabilities of a Markov process; its solution would provide a complete description of the Markov process. However it's a rather complex non-linear functional equation for which no general solution is known. Nonetheless, this equation is the hook that allows to switch from the microscopic level (SDE) to the macroscopic one.

3.2.3 Macroscopic Level

Given the mesoscopic level of representation, it is possible to obtain two macroscopic limits of CK Equation, namely the macroscopic limit in time or in space. These two are represented by the **Master Equation** and the **Fokker-Planck Equation**, respectively.

The Master Equation

Here is the derivation that allows to get the Master Equation as a limit in time of the CK Equation. To such end, in order to simplify the math, let's consider the discrete version of the CK Equation:

$$P\left(\mathbf{x}', t + \Delta t \mid \mathbf{x}_0, t_0\right) = \sum_{\mathbf{x}} P\left(\mathbf{x}', t + \Delta t \mid \mathbf{x}, t\right) P\left(\mathbf{x}, t \mid \mathbf{x}_0, t_0\right) \tag{3.31}$$

where we have set $\mathbf{x}' = x_3$, $\mathbf{x}_0 = x_1$, $t + \Delta t = t_3$ and $t_0 = t1$ in order to highlight the presence of a starting state $\mathbf{x}_0$ occurring at time t_0 and an arrival state $\mathbf{x}'$ at time $t + \Delta t$. Differently from Equation 3.30, Equation 3.31 describes a Markov process

occurring on a discrete lattice of states. By assuming Δt to be a small time interval, let's define the short time conditional probability $P(\mathbf{x}', t + \Delta t \mid \mathbf{x}, t)$, where $\mathbf{x}$ represents the intermediate state between the initial state $\mathbf{x}_0$ and the arriving one $\mathbf{x}'$.

Moreover, let $w(\mathbf{x}' \mid \mathbf{x})$ denote the density variation per unit time when transitioning from $\mathbf{x}$ to $\mathbf{x}'$, and assume that such quantity is proportional to time. Thus:

$$P(\mathbf{x}', t + \Delta t \mid \mathbf{x}, t) \approx \Delta t \times w(\mathbf{x}' \mid \mathbf{x}) \tag{3.32}$$

This relation holds if $\mathbf{x}' \neq \mathbf{x}$; in order to take into account the case when $\mathbf{x}' = \mathbf{x}$, let's define the quantity:

$$Q(\mathbf{x}) = 1 - \Delta t \sum_{\mathbf{x}' \neq \mathbf{x}} w(\mathbf{x}' \mid \mathbf{x}) \tag{3.33}$$

which represents the density transition rate in the case of no state transition. Hence:

$$P(\mathbf{x}', t + \Delta t \mid \mathbf{x}, t) \approx \Delta t \times w(\mathbf{x}' \mid \mathbf{x}) + Q(\mathbf{x})\delta_{\mathbf{x}',\mathbf{x}} \tag{3.34}$$

where $\delta_{\mathbf{x}',\mathbf{x}}$ is the Kroenecker delta, assuming value 1 *iif* $\mathbf{x}' = \mathbf{x}$ and is 0 otherwise.

By substituting Equation 3.34 into Equation 3.31, after some rearrangement, we obtain:

$$\begin{aligned} P(\mathbf{x}', t + \Delta t \mid \mathbf{x}_0, t_0) - P(\mathbf{x}', t \mid \mathbf{x}_0, t_0) = \\ \Delta t \sum_{\mathbf{x}} \left[w(\mathbf{x}' \mid \mathbf{x}) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0) - w(\mathbf{x} \mid \mathbf{x}') P(\mathbf{x}', t \mid \mathbf{x}_0, t_0) \right] \end{aligned} \tag{3.35}$$

Note that by letting $\Delta t \to 0$, le left hand side of Equation 3.35 can be written as:

$$\lim_{\Delta t \to 0} \frac{P(\mathbf{x}', t + \Delta t \mid \mathbf{x}_0, t_0) - P(\mathbf{x}', t \mid \mathbf{x}_0, t_0)}{\Delta t} = \frac{\partial P(\mathbf{x}', t \mid \mathbf{x}_0, t_0)}{\partial t} \tag{3.36}$$

i.e. it is the definition of *Partial Derivative*. Using such definition in Equation 3.35 and letting $\Delta t \to 0$, we obtain:

$$\frac{\partial P(\mathbf{x}', t \mid \mathbf{x}_0, t_0)}{\partial t} = \sum_{\mathbf{x}} \left[w(\mathbf{x}' \mid \mathbf{x}) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0) - w(\mathbf{x} \mid \mathbf{x}') P(\mathbf{x}', t \mid \mathbf{x}_0, t_0) \right] \tag{3.37}$$

We can now multiply both size of Equation 3.37 by $P(\mathbf{x}_0, t_0)$ and use the discretized version of the marginalization rule described by Equation 3.6, i.e.:

$$\sum_{\mathbf{x}_0} P(\mathbf{x}', t \mid \mathbf{x}_0, t_0) P(\mathbf{x}_0, t_0) = P(\mathbf{x}', t) \tag{3.38}$$

Finally, we obtain:

$$\frac{\partial P(\mathbf{x}', t)}{\partial t} = \sum_{\mathbf{x}} \left[w(\mathbf{x}' \mid \mathbf{x}) P(\mathbf{x}, t) \right] - \sum_{\mathbf{x}} \left[w(\mathbf{x} \mid \mathbf{x}') P(\mathbf{x}', t) \right] \tag{3.39}$$

which is the **Master Equation** describing the rate of change with respect to time of the probability $P(\mathbf{x}', t)$ of the ensemble of particles (or scanpaths) of being in state (or

position) $\mathbf{x}'$ at time t. Eventually, sums can be replaced with integrals so to recover the Master equation for continuous state space processes.

Intuitively, the Master Equation can be conceived as a balance equation: the rate of change of $P(\mathbf{x}', t)$ w.r.t. time t is given by the difference between $\sum_{\mathbf{x}} [w(\mathbf{x}' \mid \mathbf{x}) P(\mathbf{x}, t)]$, i.e. the probability of moving from any state $\mathbf{x}$ to $\mathbf{x}'$, minus $\sum_{\mathbf{x}} [w(\mathbf{x} \mid \mathbf{x}') P(\mathbf{x}', t)]$, i.e. the probability of already being in state $\mathbf{x}'$ and moving to any other state $\mathbf{x}$. In other words the variation of $P(\mathbf{x}', t)$ in time is the difference between what comes in state $\mathbf{x}'$ and what gets out of it.

The Fokker-Planck Equation

We saw how one of the macroscopic views of a stochastic process is the Master Equation; this is a stochastic *Partial Differential Equation* (PDE) with derivatives with respect to time of the PDF of an ensemble of particles.

The other way of describing macroscopically a stochastic process is through the well known **Fokker-Planck (FP) Equation**. Likewise the Master Equation, the FP Equation is a PDE too, but with derivatives with respect to both time and state space.

The FP Equation can be derived starting from the CK Equation (like the Master Equation) or from the Langevin Equation. However, the explicit derivation is not carried out here, because is out of the scope of the present thesis. The interested reader can find the full drawing in Gardiner (2011).

The FP Equation for diffusive processes, in the simple one dimensional case is:

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}[a(x,t)P(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[b(x,t)^2 P(x,t)\right] \tag{3.40}$$

Likewise the Master Equation, the FP Equation (3.40) describes the evolution of a PDF in time. However it is interesting to note how the probabilities $w(\mathbf{x}' \mid \mathbf{x})$ that in the Master Equation contained the complete statistical information of the transition between states, have been now replaced by what can be interpreted as transition moments, namely $a(x,t)$ and $b(x,t)^2$ which only include partially that information. This means that the same FP equation could be found from different expressions of the transition probabilities $w(\mathbf{x}' \mid \mathbf{x})$. This denotes the macroscopic nature of such equation.

Interestingly enough, Equation 3.40 exhibits a formal link with the Langevin Equation, right through the drift ($a(x,t)$) and the diffusion ($b(x,t)^2$) terms.

3.3 Notable Processes

As a matter of fact, Markov processes represent a viable way of reducing the mathematical complexity of stochastic processes, while being at the same time very useful in modelling natural phenomena. As a consequence, Markov processes have been widely used in the scientific literature throughout many different fields (Méndez et al., 2014).

Examples of such well-known processes are, for instance, the Wiener and the Ornstein-Uhlenbeck processes (both instances of a broader class of stochastic processes called Gaussian processes) or the Poisson process, which are discussed below.

3.3.1 Gaussian processes

A stochastic process $X(t)$ is a Gaussian process, if given a finite collection of random variables $(X(t_1), \ldots, X(t_n))$ its joint probability $P(x_1, t_1; x_2, t_2; \ldots; x_n, t_n)$ follows a multivariate Gaussian PDF, which has the form:

$$\mathrm{N}(\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{3.41}$$

Hence, any realization of a gaussian process $(X(t_1), \ldots, X(t_n)) = (x_1, \ldots, x_n)$ can be seen as a sample point from a n-dimensional multivariate gaussian:

$$(x_1, \ldots, x_n) \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma) \tag{3.42}$$

A gaussian process is completely specified by its mean $\boldsymbol{\mu}$ and covariance matrix Σ. The latter is a real, symmetric and strictly positive defined $n \times n$ matrix and captures the correlations between the random variables defining the process. If we assume that the process has 0 mean, i.e. $\boldsymbol{\mu} = \mathbf{0}$, then a gaussian process is completely specified by its second order statistics; in particular given any two random variables of the process $X(t_i)$ and $X(t_j)$, the covariance matrix measures their correlation:

$$\mathrm{Cov}\left(X\left(t_i\right) X\left(t_j\right)\right) = \left\langle X\left(t_i\right) X\left(t_j\right)\right\rangle = \Sigma_{ij} \tag{3.43}$$

If the gaussian process is uncorrelated, i.e. any random variable is independent from the others, then:

$$\left\langle X\left(t_i\right) X\left(t_j\right)\right\rangle = \delta\left(t_i - t_j\right) \tag{3.44}$$

for every i and j. Equivalently $\Sigma = \mathbb{I}$, where $\mathbb{I}$ is the identity matrix. In this case a sample from $\mathrm{N}(\mathbf{0}, \mathbb{I})$ would yield a realization of white noise (c.f.r. Section 3.1.2).

Note how equation 3.44 is the autocorrelation function of the *purely random process* (Equation 3.16), but has been employed to build the covariance matrix of a multivariate gaussian PDF; for this reason, in the context of gaussian processes the autocorrelation functions are named *covariance functions*. Covariance functions determine the shape of the covariance matrix and allow to define different kinds of gaussian processes.

Simulating Gaussian Processes

As stated earlier, simulating a gaussian process is as simple as sampling from a multivariate gaussian; in general to sample from a generic multivariate gaussian $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$ it's sufficient to be able to sample from an isotropic normal PDF $\mathbf{Z} \sim \mathbf{N}(\mathbf{0}, \mathbb{I})$, then given the *Cholesky decomposition* of the covariance matrix $\Sigma = AA^\top$, a sample from $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ can be easily obtained from:

$$\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Z} \tag{3.45}$$

Consider a gaussian process evolving in time; simulating the process for n time steps would require to define and store a $n \times n$ covariance matrix. However in the specific case of Gauss-Markov processes, Markovianity and the properties of the multivariate gaussian distribution can be exploited in order to perform efficient simulations.

Recall the properties of the multivariate gaussian distribution in the simple 2-dimensional case $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2) \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$:

1. $\mathbf{X}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \Sigma_{11})$
2. $\mathbf{X}_2 \sim \mathrm{N}(\boldsymbol{\mu}_2, \Sigma_{22})$
3. $\mathbb{E}[\mathbf{X}_2 \mid \mathbf{X}_1] = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{X}_1 - \boldsymbol{\mu}_1)$
4. $\mathrm{Var}(\mathbf{X}_2 \mid \mathbf{X}_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

Now, consider two random variables of a gaussian process indexed in two successive time steps:

$$\begin{pmatrix} X_{t_i} \\ X_{t_{i+1}} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_i \\ \mu_{i+1} \end{pmatrix}, \begin{pmatrix} \sigma_{i,i} & \sigma_{i,i+1} \\ \sigma_{i,i+1} & \sigma_{i+1,i+1} \end{pmatrix}\right) \tag{3.46}$$

where $\sigma_{i,i+1} = \mathrm{Cov}\left(X_{t_i}, X_{t_{i+1}}\right)$ and $\mu_i = \mathbb{E}[X_{t_i}]$, then the PDF of the random variable at the next time step given the current follows from the properties of the multivariate gaussian:

$$X_{t_{i+1}} \mid X_{t_i} = x_i \sim \mathrm{N}\left(\mu_i + \frac{\sigma_{i,i+1}}{\sigma i, i}(x_i - \mu_i), \sigma_{i+1,i+1} - \frac{\sigma_{i,i+1}^2}{\sigma_{i,i}}\right) \tag{3.47}$$

Note how this assumes that the RV at the next time steps only depends on the previous one (Markovianity). Following Equation 3.45 a Gauss-Markov process can thus be efficiently simulated by repeatedly sampling $Z \sim \mathrm{N}(0, 1)$ and then computing:

$$X_{t_{i+1}} = \mu_i + \frac{\sigma_{i,i+1}}{\sigma i, i}(X_{t_i} - \mu_i) + \left(\sqrt{\sigma_{i+1,i+1} - \frac{\sigma_{i,i+1}^2}{\sigma_{i,i}}}\right) Z \tag{3.48}$$

3.3.2 The Wiener process

Consider the *Langevin Equation* written as an Itô SDE described by Equation 3.23; if we set the deterministic component $a(x(t), t)$ (drift) to zero and let $b(x(t), t) = \sqrt{2D}$, where D is called the diffusion coefficient, then we obtain:

$$dx = \sqrt{2D}dW(t) \tag{3.49}$$

Equation 3.49 represents the SDE describing Brownian Motion and is generally known as the Wiener Process. Broadly speaking, if Brownian Motion is the physical phenomenon, the Wiener Process is its mathematical description at the microscopic level. It owns its name to the American mathematician and philosopher Norbert Wiener who provided a rigorous mathematical formalization of the Brownian motion, proving that the trajectory of a Brownian particle is (almost) everywhere continuous but nowhere differentiable (Wiener, 1930).

Note how by performing the Euler-Maruyama discretization of Equation 3.49:

$$x_t = x_{t-1} + \sqrt{2D}\Delta W_t = x_{t-1} + \sqrt{2D\Delta t}\xi_t = x_{t-1} + k\xi_t \tag{3.50}$$

with ξ_t sampled from a zero-mean Gaussian distribution of unit variance $N(0,1)$, Equation 3.19 is recovered. Indeed the Wiener Process can be constructed from a simple random walk by letting the time intervals to become infinitesimal.

To show this, suppose to divide the half line $[0,\infty)$ into small sub-intervals of length τ so that the first sub-interval occurs at position τ, the second at 2τ, the n-th at $t = n\tau$. Each sub-interval may be a time slot in which we toss a fair coin so that at each time step is associated a random variable Z_i ($i = 1 \dots n$) that can take values:

$$Z_i = \begin{cases} \sqrt{\tau} & \text{with probability } 0.5 \\ -\sqrt{\tau} & \text{with probability } 0.5 \end{cases}$$

Note how the coin tosses are independent with each other. The resulting stochastic process Z has $E[Z] = 0$ and $\mathrm{Var}(Z) = \tau$. Let's now define the stochastic process $W(t)$ at time $t = n\tau$ as:

$$W(t) = W(n\tau) = \sum_{i=1}^{n} Z_i \tag{3.51}$$

Since $W(t)$ is the sum of i.i.d. RVs:

$$E[W(t)] = \sum_{i=1}^{n} E[Z_i] = 0 \tag{3.52}$$

and

$$\begin{aligned} \mathrm{Var}(W(t)) &= \sum_{i=1}^{n} \mathrm{Var}(Z_i) \\ &= n\,\mathrm{Var}(Z_1) \\ &= n\tau \\ &= t \end{aligned} \tag{3.53}$$

Moreover, since the the Z_i RVs are independent, the increments of $W(t)$ are independent, too. For any fixed $t \in (0,\infty)$ as $n \to \infty$, $\tau \to 0$; by the central limit theorem, for large values of n, the difference $W(t_2) - W(t_1)$ for every $t_1 < t_2$ is close to $N(0, t_2 - t_1)$. Hence, $W(t)$ can be written as:

$$W(t) \sim N(0,t) \tag{3.54}$$

As a consequence of the Donsker's theorem (a functional extension of the central limit theorem.), as $n \to \infty$, W approaches the Wiener Process. Equation 3.54 states that the position of the stochastic trajectory defined by a Wiener Process W at time t can be described by a RV distributed as a gaussian with zero mean and variance proportional to time.

The very same conclusion can be obtained more rigorously by considering the macroscopic behaviour of the Wiener Process. In particular, consider the FP Equation

(3.40); by setting $a(x,t) = 0$ and $b(x,t) = \sqrt{2D}$ as in the SDE definition (Equation 3.49), we obtain:

$$\frac{\partial P(x,t)}{\partial t} = D\frac{\partial^2 P(x,t)}{\partial x^2} \tag{3.55}$$

which is broadly known as the *heat* or *diffusion equation*. This was the mathematical description that Einstein provided when derived the diffusion equation for Brownian particles, thus addressing the problem from the macroscopic point of view (Einstein, 1905). The solution to Equation 3.55 is given by:

$$P(x,t) = \frac{1}{\sqrt{4\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right) \tag{3.56}$$

which can be easily recognized as a gaussian PDF with zero mean and variance $\sigma^2 = 4Dt$. Hence, a particle whose microscopic dynamics is described by a Wiener Process (Equation 3.49) can by described at the macroscopic level through the *heat equation* whose solution (Equation 3.56) is a time-dependent gaussian with variance growing linearly in time.

Simulating the Wiener Process

One straightforward way of simulating a Wiener process, wold be to consider the Euler-Maruyama discretization of the SDE shown in Equation 3.49:

$$x_{i+1} = x_i + \sqrt{2D}\Delta W_i = x_i + \sqrt{2D\Delta t_i}\xi_i \tag{3.57}$$

with ξ_i sampled from a zero-mean Gaussian distribution of unit variance. However it's worth noticing how the Wiener process is a Gauss-Markov process. This means that given the covariance function, the simulation can be performed efficiently using the properties listed in Section 3.3.1. The covariance function for the Wiener process can be derived as follows: given two time instants t_1 and t_2 such that. $t_1 < t_2$,

$$\begin{aligned}
C_{WW}(t_1,t_2) &= \langle W(t_1)W(t_2)\rangle \\
&= \langle W(t_1)W(t_2) - W(t_1)^2 + W(t_1)^2\rangle \\
&= \langle W(t_1)(W(t_2) - W(t_1)) + W(t_1)^2\rangle \\
&= \langle W(t_1)(W(t_2) - W(t_1))\rangle + \langle W(t_1)^2\rangle \\
&= \underbrace{\langle W(t_1)\rangle\langle (W(t_2) - W(t_1))\rangle}_{=0} + \langle W(t_1)^2\rangle \\
&= \langle W(t_1)^2\rangle = \mathrm{Var}(W(t_1)) = t_1 \\
&= min(t_1,t_2)
\end{aligned} \tag{3.58}$$

Now consider Equation 3.48, bearing in mind that for the Wiener process $\mu_i = \langle W(t_i)\rangle = 0$ and $\sigma_{i,i+1} = \min(t_i, t_{i+1}) = t_i$, the process can be simulated by recursively computing:

$$W(t_{i+1}) = W(t_i) + \left(\sqrt{t_{i+1} - t_i}\right) Z \tag{3.59}$$

where $Z \sim \mathrm{N}(0, 1)$. Note that Equation 3.59 and Equation 3.57 are equal except for a constant.

3.3.3 The Ornstein-Uhlenbeck process

Consider again, the general form of the Langevin Equation written as an Itô SDE (Equation 3.23); we have seen that, by setting the drift term to zero the diffusion term to a constant, the resulting SDE describes the Wiener process or Brownian Motion.

The *Ornstein-Uhlenbeck* (OU) process is obtained by setting the drift $a(x(t), t) = \beta(\mu - x(t))$ and the diffusion $b(x(t), t) = D$, thus getting:

$$dx(t) = \beta(\mu - x(t))dt + DdW(t) \tag{3.60}$$

where $\beta > 0$, μ and $D > 0$ are constants. If the stochastic part of Equation 3.60 does not show any particular novelty, describing *de facto* a Wiener process, the deterministic term expresses a more interesting functional form.

Intuitively, supposing to consider only the drift term, it can be deduced that the instantaneous change of x i.e. $dx(t)$ depends on the distance of the current state $x(t)$ from the value μ. In particular, if $\mu - x(t) > 0$, then $dx(t)$ will be positive; hence, $x(t)$ will increase. The opposite holds when $\mu - x(t) < 0$. As a consequence, the drift term will force $x(t)$ to move towards μ as the process evolves. For this reason the parameter μ is often called *steady state* or *attractor* and the OU process is defined as a *mean reverting* process.

Note how the strength of the attraction towards μ is modulated by the parameter β; for bigger values of β, the difference $(\mu - x(t))$ will be magnified, therefore a faster change will occur in the direction of the steady state. On the other hand, for values of β close to zero the attraction will be weaker. Because of this property, the parameter β is called *dampening force* or *centralizing tendency*.

When adding the diffusion term, thus considering Equation 3.60 entirely, it's easy to imagine how the behaviour of a trajectory whose microscopic dynamics is described by an OU process would consist is a noisy run towards μ, the amount of "noise" being determined by the parameter D. Eventually, when the steady state is reached, the process will keep wiggling around μ. This is depicted in Figure 3.7 which shows a Monte Carlo simulation of 7 trajectories, all described by the same OU process with parameters $\beta = 1$, $\mu = 0$ and $D = 0.2$, starting at different points.

The solution of the OU process is obtained by integrating the SDE of Equation 3.60. This results in an expression for $x(t) \mid x(t-d)$ i.e. a conditional on d time units before:

$$x(t) \mid x(t-d) \sim N\left(\mu + e^{-\beta d}(x(t-d) - \mu), \frac{D^2}{2\beta}\left(1 - e^{-2\beta d}\right)\right) \tag{3.61}$$

It is immediate to verify that if we let d tend to infinity, i.e. we condition on many time units earlier, the distribution does not depend on $x(t-d)$ anymore. Hence, the initial condition of the process is forgotten. In this case the solution reads:

$$x(t) \sim N\left(\mu, \frac{D^2}{2\beta}\right) \tag{3.62}$$

Figure 3.7: *7 different realizations of the same one dimensional OU process with different initial conditions. As can be noted, despite the different starting points, the process runs towards the steady state* $\mu = 0$

In contrast to the Wiener process, whose solution is a time dependent Gaussian distribution (Equation 3.54), the OU process admits a stationary probability distribution which is a Gaussian with mean μ and standard deviation depending on the constants D and β, as shown in Equation 3.62. This fact can be appreciated qualitatively by looking again at Figure 3.7.

The n-dimensional generalization of the OU process is given by:

$$d\mathbf{x}(t) = \mathbf{B}(\boldsymbol{\mu} - \mathbf{x}(t))dt + \mathbf{D}d\mathbf{W}(t) \tag{3.63}$$

Now $\mathbf{x}(t)$ represents the position of the trajectory in a n-dimensional space, which is pulled towards the steady state represented by the n-dimensional vector $\boldsymbol{\mu}$. The adjustment to the attractor is now determined by the $n \times n$ matrix $\mathbf{B}$, while the $n \times n$ covariance matrix $\mathbf{D}$ controls the variances and covariances of the n driving white noise processes $d\mathbf{W}(t)$. In order to ensure the stability of the process (convergence to the stationary distribution) the matrix $\mathbf{B}$ is supposed to have all positive eigenvalues (Oud and Singer, 2008).

Eq. 3.63 can be explicitly written in the two dimensions simply as

$$dx(t) = b_x[\mu_x - x(t)]dt + DdW_x(t), \tag{3.64}$$
$$dy(t) = b_y[\mu_y - y(t)]dt + DdW_y(t). \tag{3.65}$$

Consider the 1-D process on the x coordinate. It is known that for $t \geq 0$, with initial value $x(0) = x_0$, the explicit solution of Eq. 3.64 writes (see e.g. Lemons (2002); Kloeden and Platen (2013)):

$$x(t) = x_0 e^{-b_x t} + \mu_x(1 - e^{-b_x t}) + D_x^2 \int_0^t e^{-b_x(t-s)} dW_x(s), \tag{3.66}$$

and analogously for the $y(t)$ process. The solution can be equivalently written as Equation 3.61 by conditioning on the initial state $x(0)$:

$$x(t) \mid x(0) \sim \mathcal{N}(\mu_x + e^{-b_x t}(x_0 - \mu_x), \gamma_x(1 - e^{-2b_x t})), \tag{3.67}$$

with $\gamma_x = \frac{D^2}{2b_x}$, so that the expected value is $\mathbb{E}[x(t)] = \mu_x + e^{-b_x t}(x_0 - \mu_x)$ and the variance ia $var(x(t)) = \gamma_x(1 - e^{-2b_x t})$. The same holds for the $y(t)$ process.

The explicit evolution of x in time between 0 and t can be obtained by numerically advancing the particle position with an update equation. This is derived by replacing t in the exact solution (Eq. 3.66) with $t' = t + \delta t$, δt time units later, and applying the initial condition $x_0 = x(t)$:

$$x(t') = x(t)e^{-b_x \delta t} + \mu_x(1 - e^{-b_x \delta t}) + \sqrt{\gamma_x(1 - e^{-2b_x \delta t})} z(t). \tag{3.68}$$

In the same way, Eq. 3.67 writes as the conditional distribution

$$x(t') \mid x(t) \sim \mathcal{N}(\mu_x + e^{-b_x \delta t}(x(t) - \mu_x), \gamma_x(1 - e^{-2b_x \delta t})). \tag{3.69}$$

Interestingly enough, Eqs. 3.68 and 3.69 can be read as solving Eq. 3.64 via Monte Carlo simulation, where a sequence of such updates with the realization of the updated position $x(t')$ at the end of each time step is used as the initial position $x(t)$ at the beginning of the next.

Eventually, Eq. 3.69 and the corresponding one for the $y(t)$ coordinate can be generalized in compact form as

$$\mathbf{x}(t') \mid \mathbf{x}(t) \sim \mathcal{N}(\boldsymbol{\mu} + e^{-\mathbf{B}\delta t}(\mathbf{x}(t) - \boldsymbol{\mu}), \boldsymbol{\Psi}), \tag{3.70}$$

which represents the general solution to Eq. 3.63, with

$\boldsymbol{\Psi} = \boldsymbol{\Gamma} - e^{-\mathbf{B}\delta t}\boldsymbol{\Gamma}e^{-\mathbf{B}'\delta t}$; $\mathbf{B}$ and $\boldsymbol{\Gamma} = \frac{D^2}{2}\mathbf{B}^{-1}$ are 2×2 matrices and $e^{-\mathbf{M}}$ is the matrix exponential.

If $\mathbf{B}$ is positive definite the OU process is stationary; intuitively, by letting $\delta t \to \infty$, then $e^{-\mathbf{B}\delta t} \to 0$ and the process has the equilibrium distribution $\mathbf{r}(t) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma})$. The assumption of stationarity implies that if the process runs for an infinitely long period of time, this equilibrium density is the density function of the visited points in the 2-dimensional space.

By examining the conditional mean vector

$$\mathbb{E}\left[\mathbf{x}(t') \mid \mathbf{x}(t)\right] = \boldsymbol{\mu} + e^{-\mathbf{B}\delta t}(\mathbf{x}(t) - \boldsymbol{\mu}),$$

the parameter $\boldsymbol{\mu}$ is the vector of expected values of the equilibrium distribution ($\delta t \to \infty$) and can thus be seen as a fixed point attractor; the matrix $\mathbf{B}$ controls the strength of the centralising tendency, which keeps the process in the vicinity of attractor $\boldsymbol{\mu}$.

The matrix $\boldsymbol{\Gamma}$ is the covariance matrix of the stationary distribution in and is part of the conditional covariance. It is a positive definite, symmetric 2×2 matrix; large variance values imply that the gaze process can go through many changes (i.e., it is very volatile), while small variances lead to smoother trajectories. As to the instantaneous variance $\boldsymbol{\Psi} = \boldsymbol{\Gamma} - e^{-\mathbf{B}\delta t}\boldsymbol{\Gamma}e^{-\mathbf{B}'\delta t}$, we can see that when the exponential part goes to 0 (i.e., a large centralizing tendency and/or time difference), the instantaneous variance converges to the variance of the stationary distribution. As the exponential part goes to 1 (i.e., small centralizing tendency and/or time difference), the conditional variance becomes very small.

Simulating an Ornstein-Uhlenbeck process

As the Wiener process, the OU process is a Gauss-Markov process, moreover it is stationary. The covariance function for the OU process is given by:

$$C_{xx}(t_1, t_2) = e^{\frac{\beta|t_1 - t_2|}{2}} \tag{3.71}$$

By substituting $\sigma_{i,i+1} = e^{\frac{\beta|t_i - t_{i+1}|}{2}}$ in Equation 3.48, for any $\mu \in \mathbb{R}$ and $D > 0$ we obtain:

$$x_{t_{i+1}} = \mu + e^{\frac{\beta|t_1 - t_2|}{2}}(x_{t_i} - \mu) + \left(\sqrt{1 - e^{\beta|t_1 - t_2|}}\right) DZ \tag{3.72}$$

Hence, an OU process can be simulated by recursively evaluating Equation 3.72 where $Z \sim N(0, 1)$.

3.3.4 The Poisson process

The Poisson process is one of the most widely used stochastic processes for modelling the times at witch certain events occur. It is a continuous time process, meaning that such events may occur at arbitrary positive times $0 < S_1 < S_2 < \cdots$, where $S_{i+1} - S_i$ is a random variable defined in $\mathbb{R}^+$. The RVs S_i are called *arrival times* or *arrival epochs* and represent the times at which some repeating phenomenon occurs. This kind of process are broadly known as *arrival processes*. An example of an arrival process is given in Figure 3.8, which shows a sequence of arrival times $(S_1, S_2, \ldots)$.

However, it's easy to note how such processes can be easily defined in two alternative ways; the first one is given by the sequence of inter-arrival times $X_1, X_2, \ldots$, which are positive RVs defined in terms of the arrival times: $X_1 = S_1$, $X_i = S_i - S_{i-1}$.

Equivalently the arrival times can be defined in terms of inter-arrival times as $S_n = \sum_{i=1}^{n} X_i$. If $X_1, X_2, \ldots$ is a sequence of i.i.d. RVs, then the corresponding arrival process is called a *renewal process*.

The other alternative to define an arrival process is through the *counting process* $N(t)$, with $t > 0$. $N(t)$ represents the number of events that have occurred up to time

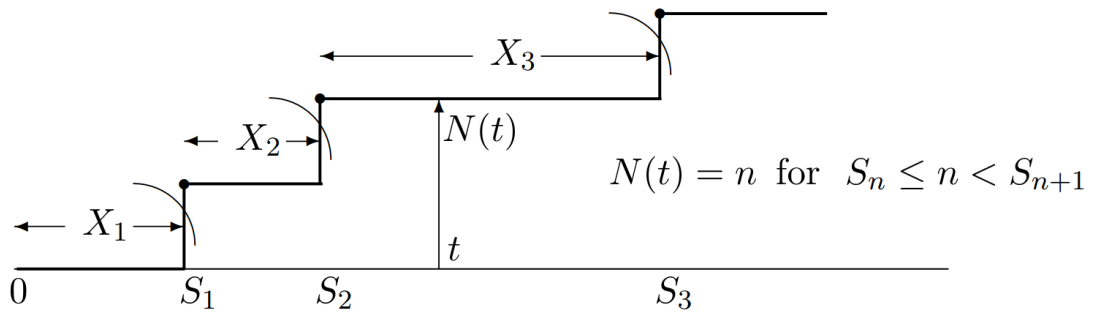


Figure 3.8: *A realization of a random sequence of arrival times, together with inter-arrival intervals and its counting process*

t. $N(0)$ is defined to be 0 with probability 1 and $N(t_2) - N(t_1)$ for every $t_2 > t_1$, is a positive and discrete random variable.

The Poisson process is most commonly viewed as a counting process. For a fixed $\lambda > 0$, the counting process $\{N(t), t \in [0, \infty)\}$ is called a *Poisson Process* with rate λ, if:

1. $N(0) = 0$
2. $N(t)$ has independent increments
3. the number of arrivals in the unitary time interval has $Poisson(\lambda)$ distribution

where $Poisson(\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$ for $k = \{1, 2, 3, \ldots\}$.

Hence, the Poisson process allows to model the number of events that occur in a specific time interval. For a generic time interval $\tau = t_2 - t_1$, the number of arrivals is given by a $Poisson(\lambda\tau)$. Consequently, the distribution of the number of arrivals in any interval depends only on the length of the interval, and not on the exact location of the interval on the real line. Therefore the Poisson process has stationary increments.

3.4 Order in apparent chaos: diffusion and Central Limit Theorems

In 1889 Sir Francis Galton, first cousin of Charles Darwin and mentor of the father of frequentist statistics Karl Pearson, published a treatise about the laws of heredity entitled *Natural Inheritance* (Galton, 1894). He formulated a theory about the inheritance of human traits such as stature and intelligence. In order to demonstrate his theories he came up with a device that he called "quinqunx" (now usually referred to as the Galton board or bean machine); it consists of a vertical board with interleaved rows of pegs into which small metal balls can be inserted through an opening at the top so that they bounce either left or right as they hit the pegs. Eventually, balls are collected into bins at the bottom of the board, and a bell shaped histogram appears.

The quinqunx was likely built about 16 years earlier and its original purpose was to give a mechanical illustration of *"the principle of the Law of Error of Dispersion"*, but Galton gave a detailed description of this device only lately in his book (Kunert et al., 2001).

The way the quinqunx was employed by Galton to deal with his theories about inheritance is out of the scope of this dissertation; still and all, his device witnesses the ubiquity of the Gaussian distribution in the statistical description of natural phenomena; Galton himself, used it as a teaching aid, e.g. for a lecture at the Royal Society in 1874.

Galton was fascinated about the fact that although the path followed by any individual ball looked completely random, when a bunch of balls are introduced into the board a nice bell shaped curve appeared into the bins at the bottom. He was charmed by the *order* of the bell curve that emerges from the *apparent chaos* of balls bouncing off of pegs in the quinqunx; this is well described by this quote from his book (Galton, 1894):

> *Order in apparent Chaos. - I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along. The tops of the marshalled row form a flowing curve of invariable proportions; and each element, as it is sorted into place, finds, as it were, a pre-ordained niche, accurately adapted to fit it.*

This order, of course, has a mathematical explanation. By recalling the construction of the Wiener process presented in the previous Section, it's easy to think about the pegs, making the ball bouncing either left or right, as a series of coin flips, and the path of the single ball as the process described by Equation 3.51. The bell-shaped distribution appears at the bottom of the Galton board for the same reason that led to the construction of the Wiener process: the *Central Limit Theorem*, proved in 1810 by Pierre-Simon Laplace. The Galton board is simply a visual demonstration of Laplace's theorem.

The wide applicability that the Gaussian distribution enjoys for the statistical description of natural phenomena is a direct consequence of the Central Limit Theorem (CLT). The key idea behind the classical CLT is the following. If we sum a large number n of RVs X_i that are:

- statistically independent
- identically distributed
- with finite and non-zero variance

the resulting probability distribution $P(S_n)$ for the sum $S_n = \sum_{i=1}^{n} X_i$ converges to a Gaussian distribution with mean $n\mu$ and variance $n\sigma^2$. In other words, as $n \to \infty$ $P(S_n) \sim N(n\mu, n\sigma^2)$.

This represents the "classical" and most famous version of the CLT; however, there exist other weaker versions that do not require the random variables X_i to be identically

distributed. According to such versions (Lyapunov CLT, or Lindeberg CLT), there may be distributions with different variances $\mathrm{Var}\,(X_i) = \sigma_i^2$. In this case, the CLT still holds if the contribution of any individual random variable to the overall variance $\sigma_n^2 = \sum_{i=1}^{n} \sigma_i^2$ is arbitrarily small for $n \to \infty$.

One of the properties that enable the classical CLT is the fact that the linear combination of independent Gaussian distributions yields again a Gaussian. Such property is called *stability* and distributions enjoying it are called *Stable distributions*. Besides the Gaussian, other functions of the exponential family share this property; these can be gathered into a broad collection of functions, which is generally referred to as the family of Lévy alpha-stable distributions, after Paul Lévy, the first mathematician to have studied it (Mandelbrot, 1960). In general, the density function of an alpha-stable random variable cannot be given in closed form. However, the characteristic function can always be given:

$$\varphi(t;\alpha,\beta,\gamma,\delta) = \exp\left(it\delta - |\gamma t|^{\alpha}(1 - i\beta\,\mathrm{sgn}(t)\Phi)\right) \tag{3.73}$$

where

$$\Phi = \begin{cases} \tan\left(\frac{\pi\alpha}{2}\right) & \alpha \neq 1 \\ -\frac{2}{\pi}\log|t| & \alpha = 1 \end{cases}$$

By computing the Fourier transform of $\varphi(t;\alpha,\beta,\gamma,\delta)$, one obtains the expression for the Lèvy stable distribution:

$$f(x,\alpha,\beta,\gamma,\delta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi(t)e^{-ixt}dt \tag{3.74}$$

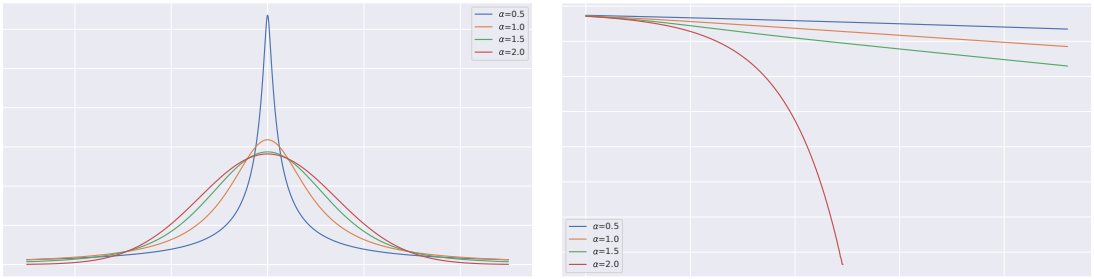
The α-stable distribution is a four-parameter family of distributions $f(x,\alpha,\beta,\gamma,\delta)$. The first parameter $\alpha \in (0,2]$ is called the *characteristic exponent*, and describes the tail of the distribution; the closer to 0 the more heavy tailed the distribution is. The $\beta \in [-1,1]$ parameter is the *skewness*, and as the name implies, specifies if the distribution is right- ($\beta > 0$) or left- ($\beta < 0$) skewed. The last two parameters are the scale, $\gamma > 0$, and the location $\delta \in \mathbb{R}$. One can think of these two as being similar to the variance and mean in the Normal distribution.

The family of alpha-stable distributions is a rich class, and includes the following distributions as sub-classes:

- Gaussian distribution $N(\mu,\sigma^2)$ is recovered from $\varphi(2,\beta,\frac{\sigma}{\sqrt{2}},\mu)$
- Cauchy distribution with scale γ and location δ is given by $\varphi(1,0,\gamma,\delta)$
- Levy distribution with scale γ and location δ is given by $\varphi(\frac{1}{2},1,\gamma,\delta)$

Figure 3.9a shows the α-stable distribution for different values of the characteristic exponent α ($\alpha = 0.5, 1, 1.5, 2$), and fixed values of the other parameters ($\beta = 0$, $\gamma = 1$, $\delta = 0$) . Notably, for $\alpha = 2$ the bell shape of the Gaussian distribution is recovered; on the other hand, as α approaches 0, the distribution becomes more picked and with fatter tails. Mandelbrot referred to such distributions as *"stable Paretian distributions"* due to the fact that the trend of the tails follows a power law or "paretian" behaviour. In

particular, the tails decrease as $|x|^{-\alpha-1}$ (Mandelbrot, 1963). This can be appreciated by looking at Figure 3.9b; it shows the Complementary Cumulative Distribution Function (CCDF) of the α-stable distribution for different values of the characteristic exponent α on a log-log plot. Notably, for $\alpha = 2$ (the Gaussian case) the tails exhibit a rapid fall-off. On the contrary, for $\alpha < 2$ the tail distribution on the log-log scale shows almost straight lines, thus denoting a power law tail behaviour. Due to this property, the α-stable distribution has undefined (or diverging) variance for $\alpha < 2$.



(a) α-*stable PDF* **(b)** α-*stable CCDF*

Figure 3.9: ***Left:*** *α-stable distribution for different values of the characteristic exponent α ($\alpha = 0.5, 1, 1.5, 2$).* ***Right:*** *α-stable Complementary Cumulative Distribution Function (CCDF) for different values of the characteristic exponent α*

The α-stable distribution plays an important role in what has been called the *Generalized Central Limit Theorem*. Such generalization to the classical CLT is due to Gnedenko and Kolmogórov (1954) and states that the sum of a number n of random variables with symmetric distributions having power-law tails (and thus infinite variance), will tend to an α-stable distribution as $n \to \infty$.

3.4.1 Normal Diffusion: Brownian Motion

In Section 3.3.2, it has been shown how the Wiener process can be seen as a limiting case of a simple random walk with positive or negative increments, provided by a coin toss. The Wiener process is thus recovered as the sum of the resulting independently sampled Bernoulli RVs; when the number of samples tend to infinity, the large sample of independent and identically distributed random displacements (with finite variance) produces the Gaussian PDF as distribution of the step length, namely $W(t_2) - W(t_1) \sim N(0, \sigma^2)$.

This construction of the Wiener process, put the accent on how classic CLT is the foundation of Brownian Motion and explains why it is so widely employed for the description of natural phenomena.

When considering the long term limit of Brownian Motion, either following the above construction (Equation 3.54), or by examining the macroscopic limit given by the FP Equation (Equation 3.56), the result is a Gaussian distribution with variance growing linearly in time.

More precisely, this fact can be further verified by inspection of the *Mean Squared Displacement (MSD)*, a measure of the deviation of the position of a trajectory with respect to a reference position over time. Intuitively, it measures the amount of "space"

that the random walker explores up to a given time instant. The MSD of a walk that starts at position x_0 at time t_0 is:

$$MSD = \left\langle |x - x_0|^2 \right\rangle \tag{3.75}$$

For Brownian Motion Einstein (1906) showed that $MSD = 2Dt$; this can be written in a more general form in terms of the *Hurst Exponent* H:

$$MSD = kt^{2H} \tag{3.76}$$

where $H = 0.5$ for Brownian Motion. The Hurst Exponent H captures the long term behaviour of a time series in terms of its correlation properties. More generally, the value of H in Equation 3.76 allows to determine the relation of the process with respect to time. For $H < 0.5$ the MSD grows sub-linearly in time; $H > 0.5$ denotes a super-linear growth, while $H = 0.5$ indicates a linear trend.

The latter behaviour is known as *Normal Diffusion* and is exhibited only by the processes obeying the rules of the classic CLT. By contrast, when classic Central Limit Theorem cannot be applied, the so called *Anomalous Diffusion* arises.

3.4.2 Anomalous Diffusion: Deviating from the CLT

Often, and in particular when dealing with the statistical description of eye movements, some of the conditions required by the classic CLT are not fulfilled (cfr Section 3.5).

In this respect, a violation of the first condition (*independence*), would lead to trajectories exhibiting long term correlations, so a movement of the random walker in a given direction would influence the probability of moving in the same direction at successive time steps. In this case, *superdiffusion* would arise. On the contrary, the process would reveal long pausing times (portion of time in which the process holds in a resting state), thus expressing *subdiffusion*.

Violating the second condition (RV are not *identically distributed*), would mean having processes with non-identical displacements. This may be gradually shorter or longer, leading to either subdiffusion or superdiffusion, respectively. These two are often collapsed into a single condition (the i.i.d. property).

A violation of the third condition would imply innovation distributions with infinite variance.

Two widely known examples of stochastic processes not ruled by classical CLT are *Fractional Brownian Motion* and *Levy Flights*.

Fractional Brownian Motion

A celebrated example of a stochastic process violating the i.i.d. condition is the **Fractional Brownian Motion** (FBM), proposed by Mandelbrot and Van Ness (1968). It is a Gaussian process whose covariance function is given by:

$$C_{xx}(t_1, t_2) = \frac{1}{2}\left(t_1^{2H} + t_2^{2H} - |t_1 - t_2|^{2H}\right) \tag{3.77}$$

where $H \in (0, 1)$ is the Hurst exponent. Observe that for $H = 0.5$ the covariance function of the Wiener process or standard Brownian Motion is recovered. Hence the

FBM represents a generalization of standard Brownian Motion obtained by allowing the Hurst parameter to differ from 0.5.

The Hurst exponent H governs all essential properties of FBM. In a nutshell, for $H \in (0, 0.5)$ the process is called *counter-persistent* i.e. if it was increasing in the past, it is more likely to decrease in the future, and vice versa. In other words, increments are negatively correlated. The opposite holds for $H \in (0.5, 1)$, the process is called *persistent*, i.e. random walk shows the tendency to continue to move in the current direction and exhibits long term memory. This properties follow from the fact the FBM is characterized by increments that for $H \neq 0.5$ are dependent, hence classical CLT does not apply here.

It follows from the above discussion that the Hurst parameter H dictates the regularity of FBM: the closer H is to 1, the smoother the process becomes. This can be appreciated by looking at Figure 3.10 showing three realization of Fractional Brownian Motion with different values of the Hurst exponent H ($H = 0.1, 0.5, 0.9$).

(a) *FBM with* $H = 0.1$ **(b)** *FBM with* $H = 0.5$ **(c)** *FBM with* $H = 0.9$

Figure 3.10: ***Left:*** *A realization of a counter-persistent FBM with* $H = 0.1$*; increments are negatively correlated and the resulting process is very irregular.* ***Center:*** *A realization of FBM with* $H = 0.5$ *(standard Brownian Motion)* ***Right:*** *A realization of a persistent FBM with* $H = 0.9$*; increments are positively correlated and the process is smoother.*

If we consider the MSD for Fractional Brownian Motion $MSD = kt^{2H}$ it is easy to see how for $H < 0.5$ the deviation of the random walk from the initial position grows sub-linearly in time (subdiffusion). On the other hand, for $H > 0.5$ the MSD evolves super-linearly (superdiffusion).

Lèvy Flights

The third condition to be met in order to apply the classical CLT is the use of RVs described by distributions having finite (and non zero) variances. Violating this condition would lead to avoid convergence to Brownian Motion in the long time limit. A way of doing so, is to employ power-law tailed distributions in the random walk steps, rather than the classical Gaussian distributed displacements. When the adopted heavy-tailed distribution is in the family of α-stable distributions, the so called *Lèvy Flights* (LFs) take place. These are stochastic processes characterized by the occurrence of extremely long jumps, whose length is described by Lèvy α-stable statistics with a power-law tail and divergence of the second moment (i.e. $\alpha < 2$).

The microscopic description of a LF is straightforward; indeed these are Markov processes with stochastic increments sampled from a α-stable distribution. The SDE describing a LF can thus be obtained from Equation 3.23 by setting $a(x, t) = 0$, and replacing the stochastic increment $dW(t) \sim \xi(t)dt$ where $\xi(t) \sim N(0, 1)$, with

$dL_\alpha(t) = \psi(t)dt$ with $\psi(t) \sim f(\psi, \alpha, \beta, \gamma, \delta)$:

$$dx(t) = b(x,t)dL_\alpha(t) \tag{3.78}$$

Notably, by setting $\alpha = 2, \gamma = \frac{1}{\sqrt{2}}, \delta = 0$ and $b(x,t) = \sqrt{2D}$ standard Brownian Motion is recovered.

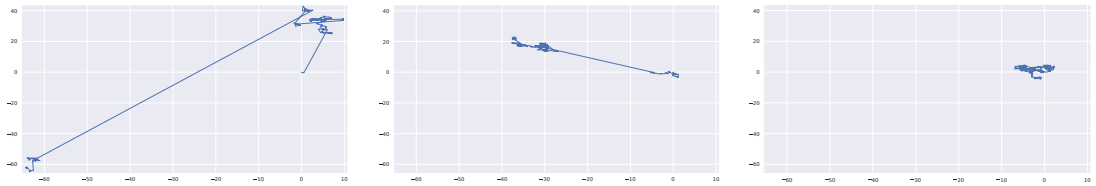
Suppose to set $b(x,t) = 1$ for simplicity, LFs can be easily simulated by discretizing Equation 3.78, thus obtaining:

$$x_{i+1} = x_i + k\psi_i \tag{3.79}$$

where $\psi_i \sim f(\psi, \alpha, \beta = 0, \gamma = 1, \delta = 0)$. The process can be straightforwardly extended in two dimensions by defining:

$$\begin{aligned} x_{i+1} &= x_i + k\psi_{x,i} \\ y_{i+1} &= y_i + k\psi_{y,i} \end{aligned} \tag{3.80}$$

Figure 3.11 depicts the simulation of 2-dimensional LFs for 500 time steps while varying the parameter $\alpha = 1, 1.5, 2$. As can be noted, the amount space explored by a Lèvy random walk is bigger for lower values of α; as α approaches 2 (Brownian Motion) the scale of the exploration is shrunk.



(a) *LF with* $\alpha = 1$ **(b)** *LF with* $\alpha = 1.5$ **(c)** *LF with* $\alpha = 0.2$

Figure 3.11: ***Left:*** *A realization of a Lèvy Flight with* $\alpha = 1$ *(Cauchy Walk)* ***Center:*** *A realization of a Lèvy Flight with* $\alpha = 1.5$ ***Right:*** *A realization of a Lèvy Flight with* $\alpha = 2$ *(standard Brownian Motion)*

This provides the notion of LFs as super-diffusive processes at an intuitive level. More rigorously, one could imagine to compute the MSD for LFs; however, the diverging second moment characterizing the α-stable distribution for $\alpha < 2$, leads to a divergent MSD. Hence, in general for LFs $MSD = \infty$.

Nevertheless, one could pursue a Monte Carlo approach by simulating many 1D LFs and computing estimates of the time varying distribution $P(x,t)$. The evolution over time of the width of such distributions can be defined empirically through, for instance, the Full Width at Half Maximum (FWHM), an expression of the extent of a function given by the difference between the two extreme values of the independent variable at which the dependent variable is equal to half of its maximum value. The FWHM is a measure of the spread of a function; if such function is the density of a Gaussian distribution it is proportional to it's standard deviation.

Figure 3.12 shows one such simulation of 1000 LF trajectories; red dots represent the FWHMs of the empirical estimate of $P(x,t)$ at fixed time instants. By looking at

the Figure it's easy to see how the growth of this "pseudo-MSD" closely follows the function $t^{\frac{1}{\alpha}}$. The latter is depicted by the black dashed line.

It's worth noticing how for $\alpha = 2$ (Brownian Motion) the increments become Gaussian and the pseudo-MSD would grows as $\sqrt{t}$; by recalling the relationship between the FWHM and the standard deviation of the Gaussian, the linear growth of the variance of the Gaussian describing BM at the macroscopic level, is easily recovered. By contrast, for $\alpha < 2$ the super-linear growth or the pseudo-MSD reveals the superduffusive nature of LFs.

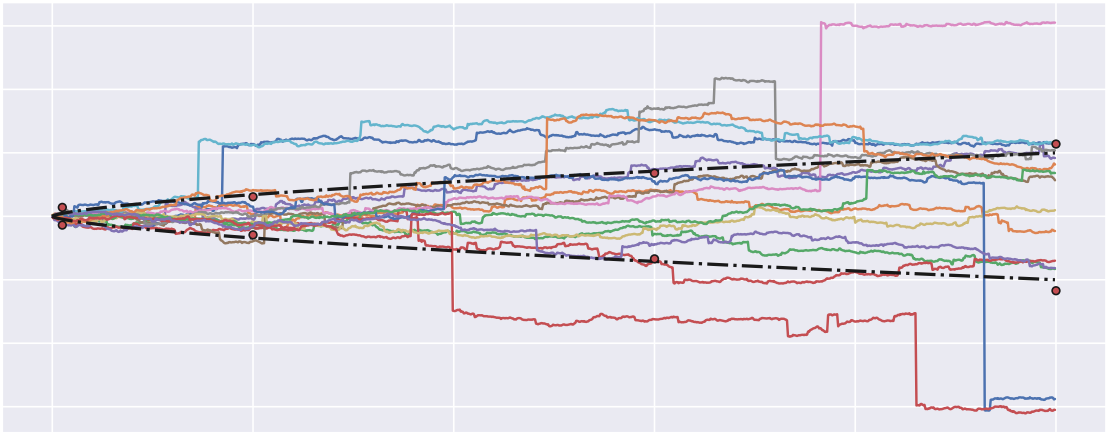


Figure 3.12: *Monte Carlo simulation of* 1000 *Lèvy Flights. The Mean Square Displacement is approximated by computing at fixed time instants* t *the Full Width at Half Maximum of the empirical distribution* $P(x,t)$ *(red dots). Dashed lines show how such "pseudo-MSD" grows as* $\sim t^{\frac{1}{\alpha}}$*, where* α *represents the characteristic exponent of the* α*-stable distribution defining the LF*

3.5 Stochastic models of eye movement and foraging

Let's consider again what Figure 3.1 represents; it suggests that each scanpath is a realization of a stochastic process. At this point one may ask what's the kind of stochastic process more apt for the stochastic modelling of eye movements.

By having a quick glimpse at the literature on the subject, one immediately realizes that when dealing with eye movements, the most appropriate stochastic processes are those for which the classical CLT is violated.

Among others, two interesting examples are the modelling of fixational eye movements by Engbert (2006); Engbert et al. (2011); Makarava et al. (2012); Engbert and Kliegl (2004) and the study of saccades amplitude by Brockmann and Geisel (2000).

3.5.1 Fixational eye movements as fractional Brownian motion

Engbert and colleagues analyzed the random walk behaviour of fixational eye movements (FEMs) by describing it via the mathematical formalism of fractional Brownian Motion.

FEMs can be defined as small involuntary eye movements that exhibit an erratic trajectory or random walk. There are two important types of FEMs that have been called

physiological drift or *tremor* and *microsaccades*. The former is a low-velocity movement, while the latter are rapid small-amplitude movements, which typically occur at a rate of one to two per second.

Engbert and Kliegl (2004) studied the correlation across time of FEMs as well as their degree of persistence by computing the empirical Mean Squared Displacement and estimating the their Hurst exponent H. In a similar vein, Makarava et al. (2012) carried out a more detailed analysis relying on Bayesian methods for the estimation of H.

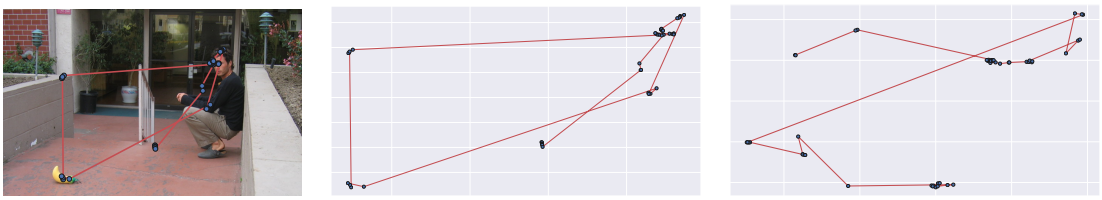
Engbert and Kliegl (2004) found that when considering a short time scale (2 to 20 ms) the random walk is persistent ($H > 0.5$); on the contrary, on larger time scales (between 100 ms and 400 ms), random walks exhibit a non-persistent behaviour ($H < 0.5$). Hence, in the short time period, fixational eye movements exhibit positive correlation between successive increments, while in the larger one, tend do be negatively correlated. According to such analysis, the latter conduct is caused specifically by the *microsaccades* that act as error-correcting movements that "balance" the diffusive carriage of the *physiological drift*.

This is psychologically plausible, because persistent behavior increases retinal image shifts, which contribute to the prevention of perceptual fading. However, a superdiffusive behaviour ($H > 0.5$) reoccurring for long time would lead to loosing the desired focus of attention (FoA). The non-persistent behaviour ($H < 0.5$) on a longer time period operates in order to keep the current fixation point. Crucially, the negatively correlated increments arising on the larger time scale, serve to avoid such disalignment (Engbert and Kliegl, 2004).

More recently Engbert et al. (2011) proposed a model of FEMs incorporating self-avoidence. In particular FEMs are considered as realizations of Self Avoiding Random Walks (i.e. random walks that keep track of the previously visited positions in order to avoid coming upon the same region twice) confined in a potential.

3.5.2 Saccades as Lèvy Flights

Consider Figure 3.13a; it shows an image and a scanpath recorded from a subject while attending at it. Figure 3.13c shows a random realization of a two dimensional Lèvy Flight. By comparing Figure 3.13b with Figure 3.13c, a pronounced similarity in terms of shape between the two trajectories should be noticeable, a least in a qualitative sense.



(a) *A sample data from the dataset Cerf et al. (2008)*

(b) *The recorded scanpath*

(c) *A realization of a 2D Lèvy Flight*

Figure 3.13: *Qualitative comparison of scanpaths and Lèvy Flights*

Brockmann and Geisel (2000) provided empirical support for such observation by assuming a power-law dependence in the tail of the saccades amplitude distribution,

thus establishing a relationship between eye movements and Lèvy Flights.

They relied their investigation on the simulation of artificial scanpaths through random walks whose step length could be sampled from either a Gaussian or a Cauchy distribution. In order to obtain simulated eye movements resembling those of humans while free viewing an image, they computed an empirical salience field from the spatial distribution of the fixations made by observers throughout the scene. This was then used to constrain the random walk (either Gaussian or Cauchian).

To verify the assumption that scanpath step length follows a power-law distribution, they collected saccadic magnitudes as measured from real subjects while scanning natural scenes, and gave evidence that their distribution's tail on a log-log plot is well fitted by a straight line. Hence, they conclude that saccades can be conceived as Levy Flights.

Moreover, they showed how employing a Cauchy Flight as saccadic model, would require a much lower time to scan the entire scene if compared with a random process with Gaussian increments. This is not surprising from a strictly mathematical standpoint: the MSD of LFs grows super-linearly, consequently they exhibit super-diffusive behaviour. However, if considering a broader perspective, such claim poses the problem of eye guidance under a new light.

Indeed, one of the findings of the work of Brockmann and Geisel (2000) is that the visual system minimizes the time needed to scan the entire visual space. Hence, they argue that our oculomotor system may have evolved in order to perform optimally, Lèvy Flights being the means ensuring such optimality requirements.

Crucially, such view allows to connect eye guidance modelling to the theories of *foraging animals* in the ecology literature, by means of what has been referred to as the **Lèvy Flight foraging hypothesis** (c.f.r. Section 3.5.3). Stated differently, one could think of the eye (or the brain modules controlling the eye behaviour) as a forager searching for valuable information (preys) in a given (and possibly time varying) scene (foraging landscape).

The forthcoming section, provides a brief introduction of those topics concerning foraging theory that are deemed relevant for the modelling of attentive eye guidance.

3.5.3 The Foraging Perspective

Foraging theory is the branch of behavioural ecology that studies the the foraging behavior of animals, i.e. the way they search for wild food resources. Indeed, animals must move in order to perform a number of fundamental tasks (eating, mating, escaping predators, etc...). The goal of foraging theory, or more specifically of the field of *movement ecology*, is to understand how living organisms move, i.e. the typical patterns and statistical properties of the trajectories describing their displacements. It is a very broad and multidisciplinary field, involving research areas (among others) like stochastic processes and anomalous diffusion.

One of the eminent ideas in behavioural ecology is the *Optimal Foraging Theory* (OFT). It was initially proposed by Emlen (1966) and MacArthur and Pianka (1966) and states that the mechanisms driving foraging organisms have been naturally selected during evolutionary time in order to maximize the energy intake. A large body of theoretical work grew in an attempt to deal with the multitude of determinant factors and in order to identify the relevant parameters involved in such optimization. Two

important examples are the *Lèvy Flight Foraging* (LFF) *Hypothesys* (Viswanathan et al., 1999, 2008) and the *Marginal Value Theorem* (MVT) (Charnov, 1976).

It turns out that many concepts of foraging theory and movement ecology successfully apply to human movement behaviour; indeed, even humans engage many forms of foraging by collecting resources from the surrounding world. For instance, Gonzalez et al. (2008) studied the movements of humans by recording GPS tracks via mobile phones; they found that after compensating for the variability among individuals, the data collapse onto an exponentially truncated Lèvy flight with characteristic exponent $\alpha = 0.75$, consistent with super-diffusion. Similarly Brockmann et al. (2006) studied human's movements by tracking the the circulation of dollar bills; they found a power law distribution of travel distances consistent with a Lèvy flight pattern of movement with $\alpha = 0.59$. Because dollar bills move only when carried by people, they conclude that the movement of people is super-diffusive.

Notably, eye movements analysis has been carried out in terms of foraging, too; aside from the already cited adoption of LFs to describe saccades (Brockmann and Geisel, 2000), Wolfe (2013) and Cain et al. (2012) examined the human's visual behaviour under such lenses, performing experiments in order to test the predictive capabilities of the MVT for visual search tasks.

In the following paragraphs of the present Chapter a quick description of such concepts (Lèvy Flight Foraging Hypothesis and Marginal Value Theorem) will be provided together with an account of their applicability to the eye movement modelling problem.

Lèvy flight foraging hypothesis

The Lèvy Flight foraging hypothesis states that since Lèvy Flights maximize the amount of space covered in a fixed time period, they optimize random searches. Hence, natural selection should have led to adaptations for Lèvy flight foraging. This is due to their super-diffusive nature arising from the power-law decay of the tails of the step length distribution.

The LFF hypothesis holds under specific but common circumstances; in particular it is assumed that the forager is engaged in a so called *non-destructive random search*, i.e. the target are randomly distributed and regenerates after some (very) short time and that these are distributed sparsely (in patches) throughout the environment.

In their seminal work, Viswanathan et al. (1999) analyzed the efficiency of a Lèvy forager for different values of the characteristic exponent $0 \leq \alpha \leq 2$ of the associated α-stable distribution. They showed how in the two extreme cases ($\alpha = 0$ and $\alpha = 2$) the forager acted sub-optimally, but for two different reasons. In the first case ($\alpha = 0$) the forager performs *ballistic movements*, this means that it chooses a random direction of movement and then moves on a straight line until a prey is encountered. This means that, eventually, closer preys are discarded.

In the opposite case ($\alpha = 2$) Gaussian innovations are preformed, the forager will thus perform Brownian motion; it will bias the search towards the closer target because it will come within the range of the bell-shaped region of the propagator before the more distant target. The linear growth of the MSD of BM makes the Gaussian propagator not very efficient (c.f.r. Figure 3.11). Moreover, this way of behaving, leads to returning often to the already visited sites (a process called *oversampling*).

The compromise is represented by Lèvy Flights with $0 < \alpha < 2$; in particular, it

has been shown (Viswanathan et al., 1999) that the optimal value of the Lèvy Flight exponent which maximizes the search efficiency of the forager is $\alpha = 1$ (Cauchy Flight). Besides such empirical findings, these results find a theoretical justification in the fact that for $\alpha = 1$, LFs reach the largest possible Hurst exponent $H = 1$, thus maximizing the super-diffusivity of the walk (Viswanathan et al., 2011).

Nonetheless, it's worth remarking that such optimal behaviour persists as long as the above conditions are met. For instance it has been shown how in case of *destructive random search*, the ballistic Lèvy searcher is the one acting optimally (Viswanathan et al., 2011). Similarly, when dealing with high-density-prey environments (targets are not patchily distributed), Lèvy Flight searches are not distinguishable from Brownian ones in terms of foraging efficiency (Viswanathan et al., 2011).

Criticism to the LLF Hypothesys

Despite the large success of the LFF hypothesis, the general applicability of such theory is still the subject of some controversy (Benhamou, 2007; Edwards et al., 2007). In fact, many of the observed patterns that are attributed to Lèvy processes can be generated by a simpler composite random walk process where the turning behaviour is spatially dependent (Codling et al., 2008; Benhamou, 2007; Bénichou et al., 2006).

In particular Benhamou (2007) argued about the fact that even if the observed patterns of movement may resemble those of LFs, (step length frequency distribution well fitted by a straight line in a log-log plot) the real generating process is not necessarily a LF process. The author shows how Composite Brownian Walks (CBW) generates search patterns that mimic those generated by LFs and that under specific circumstance are more efficient than LFs (Benhamou, 2007).

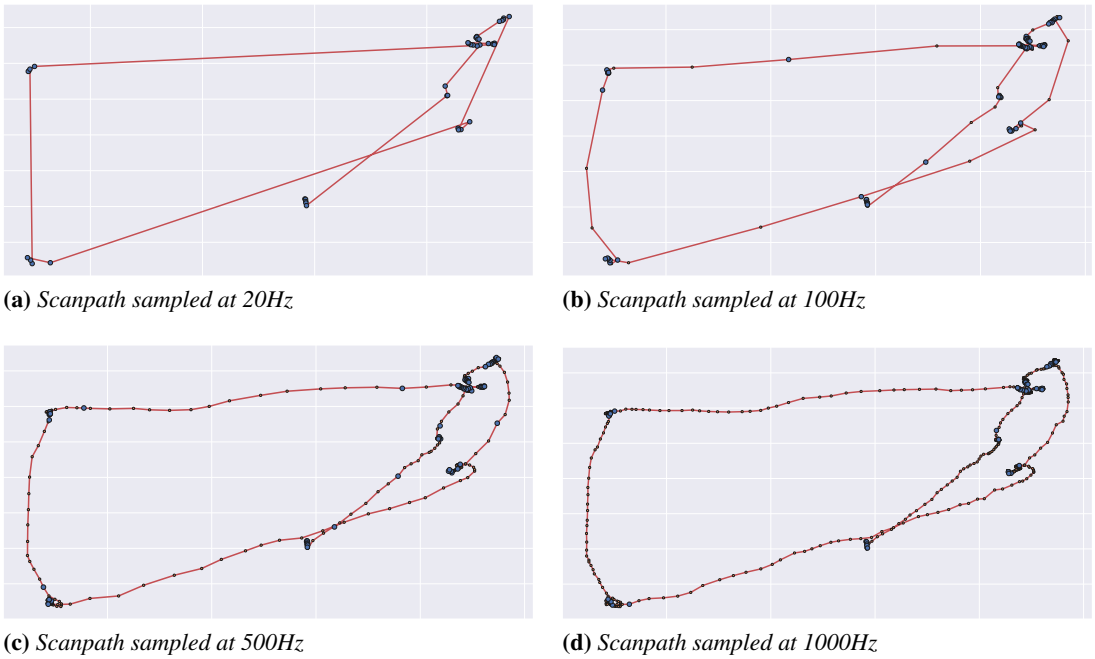
CBW are obtained as a mixture of two "classical" random walks, i.e. stochastic processes characterized by a step length distribution whose variance is finite, which act jointly to mimic the intensive and extensive search of foraging animals. One of such processes is in charge of producing frequent and exponentially distributed steps with relatively small mean, while the other produces more sporadic exponentially distributed steps with relatively large mean.

Benhamou (2007) shows how simulations from CBW may be erroneously rated as LFs. This highlights the crucial difference between the observed patterns and the inference of the data generating model and directs attention to the methods used to perform such deductions.

Following the ideas of Benhamou (2007) one could argue about the plausibility of considering the eye movements (the alternating of fixations and saccades) as generated from a mixture of simpler random walks, rather that LFs. A more rigorous account of this idea will be provided in the next Chapter. For now, it must suffice to consider the following pictorial representation.

Consider again the recorded scanpath shown in Figure 3.13 which is reported here in Figure 3.14a for the reader's convenience. Figure 3.14 shows the same scanpath at different sampling rates. The lower sampling rates are obtained via sub-sampling the original one sampled at 1000Hz. It can be seen how low sampling rates ($\sim$ 20Hz) are not able to capture the rapid eye movements like saccades, hence these seem to be ballistic displacements and the overall shape resembles a Lèvy Flight.

At finer grains (sampling rate ≥ 100), saccades are better represented. It's worth



(a) *Scanpath sampled at 20Hz*

(b) *Scanpath sampled at 100Hz*

(c) *Scanpath sampled at 500Hz*

(d) *Scanpath sampled at 1000Hz*

Figure 3.14: *The same scanpath at different sampling rates. Lower sampling rates are obtained via sub-sampling*

noticing how these are not straight lines but exhibit some curvature and randomness (c.f.r. Figure 3.15), hence these can be conceived as biased random walks towards the arriving patch. The same observation holds for fixations, which can be seen as stochastic processes with an attractor represented by the center of the patch (Figure 3.15).

In the following Section, we will build upon this idea by proposing an Ornstein-Uhlenbeck (OU) model with switching parameters as a mechanistic model of eye movements. This will allow to move from feeding (fixation) to relocation (saccades), and *vice versa* according to a switching signal defined as the output of a decision making process. The latter will be modelled, again, in terms of foraging theory by considering fixation duration as the time spent by a stochastic forager inside a patch. This can be modelled and predicted by the celebrated Marginal Value Theorem (MVT), which is described in the following.

The Marginal Value Theorem

Lèvy Flights (or CBWs) allow to model the spatial properties of the foraging path (or the scanpath), i.e. the shape of the exploration performed by the forager (observer) in a patchy environment. Crucially, they do not account for the *patch handling*, i.e. the modelling of the amount of time spent inside a patch before deciding to move to the next one.

In essence, this is a decision making process; at each time instant the forager has to decide if it's worth keeping exploit the current patch or it's time to move to next one. Indeed, patch depletion makes the encounter of preys increasingly rare, as a result the

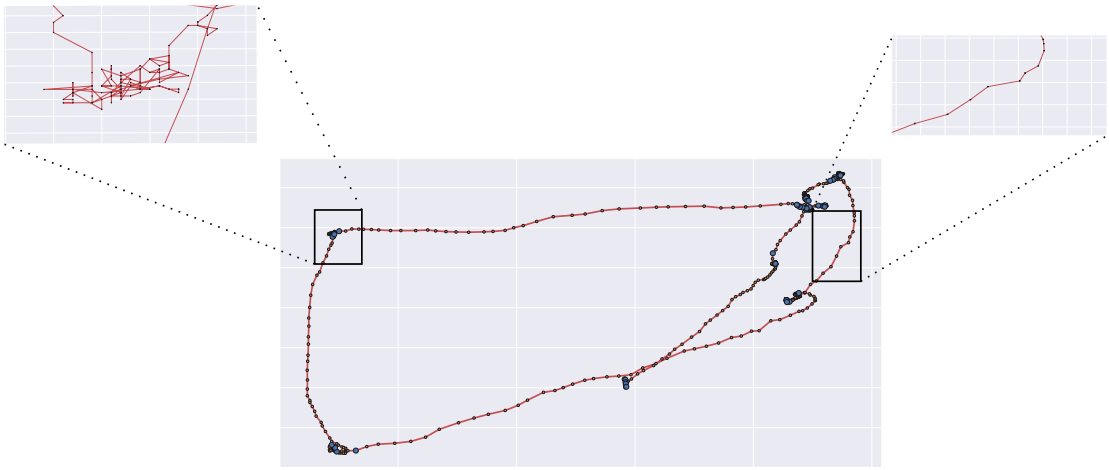


Figure 3.15: *A zoom-in on fixations and saccades as recorded by high frequency eye-trackers*

energy net intake at some point is not maximized anymore. On the other hand, leaving a patch entails some costs (typically the energy expenditure of the travel towards the next patch).

This *exploration/exploitation* dilemma, has been addressed in many Optimal Foraging Theories (Stephens, 1986); one of the most famous is the Charnov's Marginal Value Theorem (MVT) (Charnov, 1976). In a nutshell, it states that a forager moving in an environment with patchily distributed resources (separated by areas with no resources), should leave the current patch when the marginal rate of food intake drops to the long-term average rate of food gain across the patches in the environment.

Due to the resource-free space, animals must spend time traveling between patches, hence the MVT can be seen as an optimality model balancing energy gain and consumption.

The MVT assumes that exist a number n of patches, that may be differentiated for their "quality" (number of preys, quality of the preys etc..). Each patch is characterized by a *gain function*. This specifies the expected energy gain for that specific patch at time t and it is assumed to be a well-defined, continuous, deterministic and negatively accelerated function (Stephens, 1986). This means that the rate of increase of energy intake decreases as time increases. In other words a proper gain function should exhibit a steep initial slope that become progressively flatter in such a way that the patch depletion is taken into account.

Charnov (1976) showed that in order to maximize the average rate of energy intake, the forager should choose the patch residence time so that the *marginal* rate of energy gain function at the time of leaving, equals the long-term average rate of energy gain of the habitat. Here "marginal rate" is a statement inherited from economics meaning "derivative" and is responsible for the name of the Charnov's Theorem.

This solution is usually depicted graphically as in Figure 3.16. The graph shows the energy gain function associated to two types of patches A and B which are classified as "rich" and "poor", respectively. By assuming that the travel time to the patches of the environment is constant (dashed lines start at the same point), the optimal residence time or giving-up time is found by drawing the tangent to the energy gain function and projecting the point found onto the x axis. It's worth noticing how for richer patches,

the MVT predicts longer patch residence times.

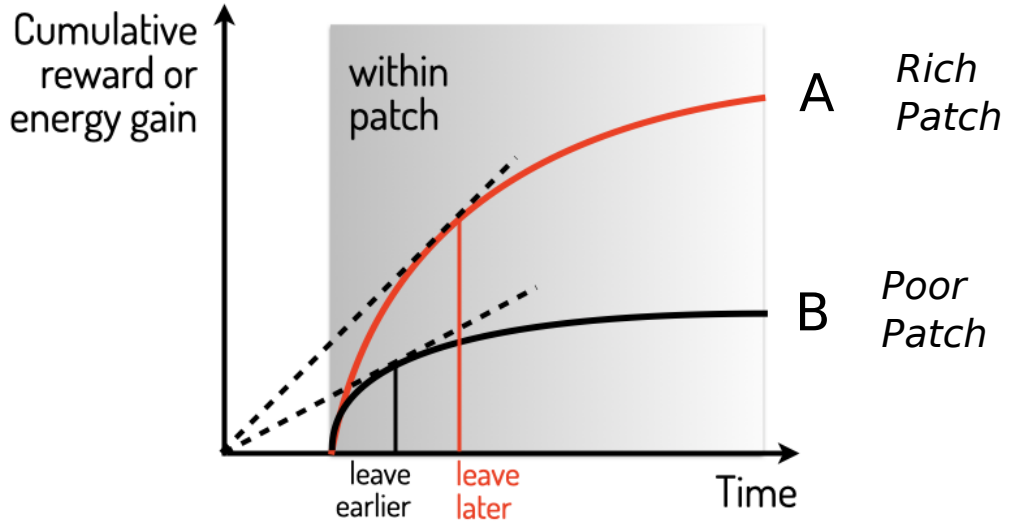


Figure 3.16: *Graphical depiction of the Marginal Value Theorem.*

Pushing ahead the simile between the animal's foraging paths and the eye movement recorded from human observers, we ask if the principles advocated by the Marginal Value Theorem can be applied to describe the patch leaving times in human visual behaviour. Putting the question straight, can MVT predict the duration of fixational eye movements?

The MVT has been previously employed by Wolfe (2013) to verify if it could predict human patch leaving behaviour in visual search experiments. It has been shown that as long as the patches are considered roughly identical, the MVT successfully predicts patch leaving times; on the contrary, as experiments start to exhibit more complex structure (different patch quality, prevent access to some information), human behaviour seems to depart from Charnov's Theorem.

In a similar vein, Cain et al. (2012) proposed an ecological optimal foraging model to quantify human strategies in multiple-target search. They employed an extended version of the MVT, the potential value theorem (McNamara, 1982) and showed that individuals searched longer when they expected more targets to be present.

In the next Chapter, we build upon such ideas and propose a full computational model of attentive eye guidance grounded on the principles of OFT. Rather than constraining the experiments on visual search specific tasks, we consider the more general free viewing condition in which observers can freely scan the scene. More specifically, we examine the human visual attentive behaviour while watching videos displaying conversations between people. Under such circumstances, the foraging eye operates in a time varying environment (video) in which patches are represented by small regions of the stimuli containing objects of interest.

The decisions about which patch to choose at any given time and how much time to spend inside each patch are demanded to a stochastic optimal foraging model relying on the MVT. This allows to solve the decision making problem associated the prediction of the current fixation duration.

The microscopic dynamics of the eye movements will be defined as a switching Ornstein-Uhlenbeck process whose switching signal is provided by the aforementioned decision making mechanism.

3.6 Summary

This Chapter provided the theoretical background of stochastic processes and diffusion, while introducing at the same time the foraging perspective on eye movements. In particular, it has been shown how gaze behaviour can be assimilated to that of foraging animals. Hence, we conjecture that principles of optimal foraging theory (OFT) can be employed to predict human overt attentive behaviour.

CHAPTER 4

A model of gaze deployment to audio-visual cues of social interaction

CONSIDER a clip displaying social interactions, in particular a conversational clip (audio and video): the chief concern of this Chapter is to model the deployment of attention through gaze by a human subject who is viewing and listening to the clip. When humans are immersed in realistic, ecological situations that involve other humans, attention deployment strives for monitoring the behaviour, intentions and emotions of others even in the absence of a given external task (Foulsham et al., 2010).

Under such circumstances, the internal goal of the perceiver is to control attention so to maximize the implicit reward in focusing signals that bear social value (Anderson, 2013).

Despite of experimental corroboration gained for such tendencies, their general modelling is far from evident. Indeed, in order to put into work the mechanisms of selection, integration and sampling underlying the multifaceted phenomenon of attention, sensory systems have to master the flood of multimodal events (e.g., visual and audiovisual) captured in the external world.

Why should this research problem be relevant beyond its merits?

One straightforward reason lies in the classic data mining hurdle. YouTube, Twitch, Facebook Live contain myriads of such clips and dedicated channels (Truong and Agrawala, 2019; Pires and Simon, 2015). Also, large-scale multimodal data conveying social interactions from non-laboratory settings are being increasingly employed to analyse behaviors, emotions, and interactions in real-life situations (Nassauer and Legewie, 2019).

It goes without saying, the processing of large spatio-temporal data from multiple

media in different contexts is a mind-blowing engineering challenge: spotting sharable highlights, capturing socially relevant events, generate value-based summaries to facilitate browsing and skimming. All such problems call for an ability that is germane to the successful performance of any cognitive task: the ability to predict and to select where the most meaningful and task-relevant information is to be found in the sensory input.

A less evident, albeit earnest need takes root in the challenge of "subject's mining": the computational inference of subject's traits, or expertise, or even expectations from his/her attentive behaviour. Much can be gained indeed by analysing the "mind's eye" conduct of a subject who scrutinizes and forages on the behaviour of other subjects involved in social interactions (Shic et al., 2007; Staab, 2014; Grossman et al., 2019; Jording et al., 2019; Guy et al., 2019).

In a nutshell, the research problem addressed here is relevant beyond its peculiar interest because it complies with a quest for parsimony. Under a variety of circumstances, what *prima facie* might come across as a conundrum of diverse engineering problems, boils down to the modelling of one and only skill: the effective deployment of attention that organisms have evolved to promote survival and well-being.

4.1 Problem statement and challenges

Throughout our lives, we are bond to unfalteringly sample the environment. Moment-by-moment we strive to answer the question: *Where to look next?* Attention guides our gaze to the appropriate location of the scene and holds it in that location for the deserved amount of time given current processing demands (Henderson, 2017).

In doing so, like other animals with as diverse evolutionary backgrounds, we exhibit a consistent pattern of eye movements. To illustrate at the finest "resolution scale" the signature of gaze dynamics, Fig. 4.1 plots the raw data recording of one subject's gaze. The trajectory of gaze is shown as unfolding in time on an excerpt of subsequent frames: large relocations are followed by local clustering of gaze points.

The given tasks or goals determine by and large such pattern (Henderson, 2017). Yet and cogent for the work described here, the pattern is not the unconcerned result of a disembodied process. Nor are the given task and the stimuli properties the only constraints to the perceiver. Subject's gut and feelings matter too: in our daily life we keenly move our gaze to gauge and collect visual information that includes social information, such as others' emotions and intentions (Shepherd and Platt, 2007; Guy et al., 2019).

The implicational converse of this state of affairs is that the dynamic pattern springing from this lifelong sampling endeavour provides information about plans, goals, interests, and probable sources of rewards; even expectations about future events (Kowler, 2011; Henderson, 2017), personality and social traits .

In this perspective, in the conversational setting, Foulsham et al. (2010) have shown that observers spend the majority of time looking at the people in the videos, markedly at their eyes and faces, and that gaze fixations are temporally coupled to the person who was talking at any one time. This is not surprising. Visually-mediated social interactions are not exclusive to humans, and have played a significant role very early in the primate lineage: selective pressure is likely to have promoted convergent evolution

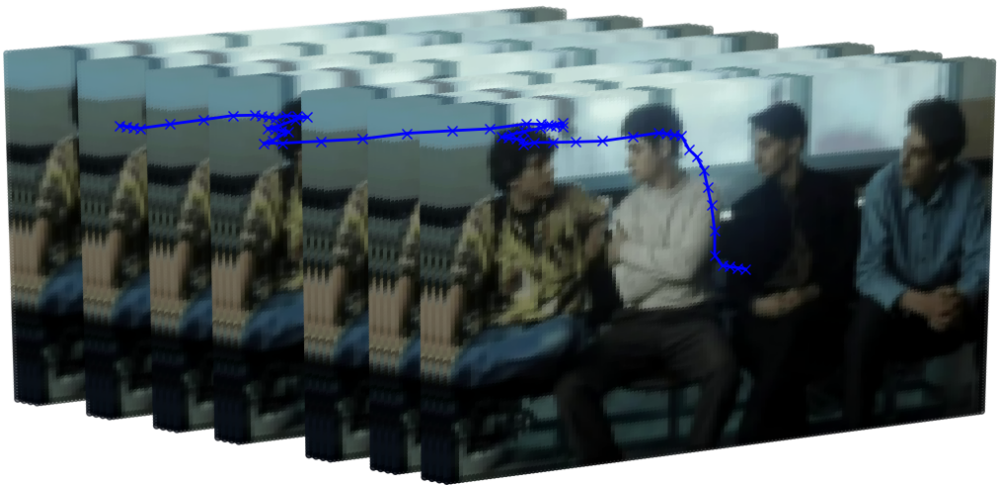


Figure 4.1: *Gaze deployment recorded from a human subject who is viewing and listening to a conversational clip. Gaze position in time is rendered by overlapping the raw data recorded along an eye-tracking session on a representative excerpt of video frames. The trajectory unfolding in time is characterised by area-concentrated phases that alternate with large distance relocations between regions attracting attention*

of social gazing abilities for social group-living animals (Shepherd and Platt, 2007).

Modelling attention in such case entails taking into account the value of social cues. This, in turn, raises the question of whether it is feasible to mine from behavioural data the implicit value of multimodal cues that drives observer's motivation.

Even prior to such urgent quest, the audio-visual nature of these stimuli brings forward the challenge of how gaze is to be guided in the context of multimodal perception (audio and visual). As discussed in Section 4.2, limited work has been devoted to eye guidance in a multimodal setting.

4.1.1 Our approach

The key intuition can be easily grasped at a glance by going back to Fig. 4.1. The trajectory of gaze unfolding in time can be best described, at the phenomenological level, as one kind of biased random walk that takes place at different scales: the fine scale of area-concentrated phases within valuable "information patches" (*exploitation*) that alternates with the coarse scale of large distance relocations between patches (*exploration*), whatever the precise rules that control them.

Thus, the portrait of Fig. 4.1 boils down our chief research problem to two crucial questions: *What* defines a valuable patch? *How* is gaze guided within and between patches?

We formalize the above intuition in a model for eye guidance that is liable to account for the characteristics of the behaviour depicted in Fig. 4.1. Namely, we consider gaze trajectories as traced by a composite forager, chasing up resources that are patchily distributed. A composite forager is one capable of switching the scale of the foraging walk from within-patch exploitation to large between-patch relocations or vice versa (Viswanathan et al., 2011). In our case, the forager is a stochastic one, and either regime - exploitation or exploration - is accomplished via a biased Brownian walk, precisely an

Ornstein-Uhlenbeck (OU) process, tuned at the appropriate scale. Crucially, the reformulation of attention in terms of foraging theory goes beyond the informing metaphor. There is substantive evidence that what was once foraging for tangible resources in a physical space became, over evolutionary time, foraging in cognitive space for information related to those resources (Hills, 2006). Such adaptations play a fundamental role in goal-directed deployment of visual attention (Wolfe, 2013).

The bias is provided by the audio-visual patches that moment-by-moment appear relevant (rewarding) within the multimodal landscape. The idea of exploiting the foraging framework has gained currency in the attention literature (cfr. Table 4.1), reckoned more than an informing metaphor (Wolfe, 2013).

Table 4.1: *Relationship between Multimodal Attention and Foraging*

Audio-visual attentive processing	**Patchy landscape foraging**
Perceiver	Forager
Perceiver's gaze shift	Forager's relocation
Audio-visual object/event	Patch
Audio-visual object/event selection	Patch choice
Deploying attention to object/event	Patch handling
Disengaging from object/event	Patch leave or giving-up

Technically, as depicted in Fig. 4.2, model input is represented by the audio-visual stream together with eye-tracking data. We exploit the publicly available dataset presented in Xu et al. (2018), who gathered data of eye-tracked subjects attending to conversational clips.

At the pre-attentive stage, inference is performed to obtain dynamic value-driven priority maps resulting from the competition of visual and audio-visual events occurring in the scene. Their dynamics integrates the observer's current selection goals, selection history, and the physical salience of the items competing for attention. The free-viewing task given to subjects allows for dynamically inferring the history of their "internal" selection goals as captured by the resulting attentive gaze behaviour.

From priority maps a number of attractors are sampled in the form of value-based patches suitable to bias the forager's walk. The attentive stage involves trading between local patch exploitation and landscape exploration through relocations across patches. This is achieved by switching the OU process at different scales. The trading rules stem from stochastic approaches to optimal foraging theory.

4.2 Background and rationale

We proceed now to set up a minimal formalism needed to outline the necessary background to the work presented here and to compare with the state-of-the-art.

Early studies on gaze behaviour and attention (James, 1890; Yarbus, 1967) made clear that in this matter three factors are to be taken into account: the task or goal $\mathcal{G}$, the stimuli $\mathcal{S}$, and the perceiver $\mathcal{O}$.

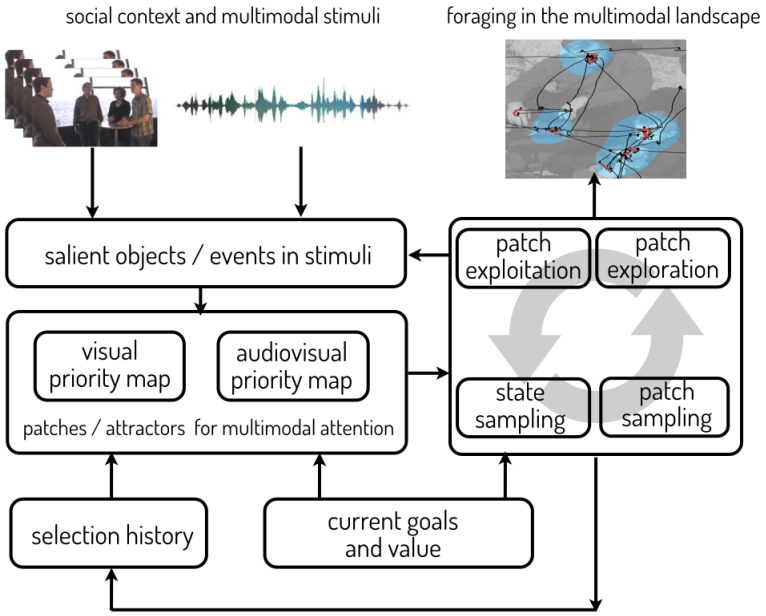


Figure 4.2: *Gaze deployment as foraging in a multimodal landscape. Model input is represented by multimodal stimuli that convey social content; the output is represented by a composite (local/global) foraging walk. Value-based patches are sampled from priority maps and integrate different sources of selection bias in a socially valuable context. The audio-visual scene social content drives perceiver's (internal) value that, in turn, guides the sampling of relevant patches. The perceiver's gaze continuously switches between local patch exploitation and between-patch global relocation. Gaze dynamics is that of a spatial Ornstein-Uhlenbeck process, which is performed at two different scales, local and global.*

Overt attention deployment as instantiated through the unfolding of gaze shifts involves two main processes: i) perception, by which $\mathcal{O}$ processes sensory information and makes inferences to set up a representation $\mathcal{W}$ capturing salient aspects of the world; ii) action $\mathcal{A}$, by which $\mathcal{O}$ chooses how to sample the world to obtain useful sensory information.

The perceptual process can be formalized in terms of an ideal perceiver model which makes task-relevant inferences. The perceiver $\mathcal{O}$ uses the sensory input $\mathcal{S}$ (visual or audio-visual, for example) together with a knowledge of the properties of the task $\mathcal{G}$ and the world, as well as features of the sensors at hand. The process of selecting an action uses both the observer's inferences and knowledge of the goal $\mathcal{G}$ to determine the next movement, i.e. where to orient the eyes. Action execution leads to new sensory input $\mathcal{S}'$. This closes the active sensing loop of perception and action.

In brief, consider time instants $t < t'$, where $t' - t = \delta t$ is an arbitrary time step. Assume that at time t the perceiver's gaze centers the focus of attention (FOA) at location $\mathbf{r}_F(t)$.

Then, the goal-driven action/perception cycle performed by $\mathcal{O}$ boils down to the iteration of the following steps. Under goal $\mathcal{G}$ and current sensory input $\mathcal{S}(t)$

Step 1: Infer the current perception of the world $\mathcal{S}(t) \rightarrow \mathcal{W}(t)$ when gazing at $\mathbf{r}_F(t)$

Step 2: Sample the appropriate motor action/decision $\mathcal{A}(t)$ depending on $\mathcal{W}(t)$;

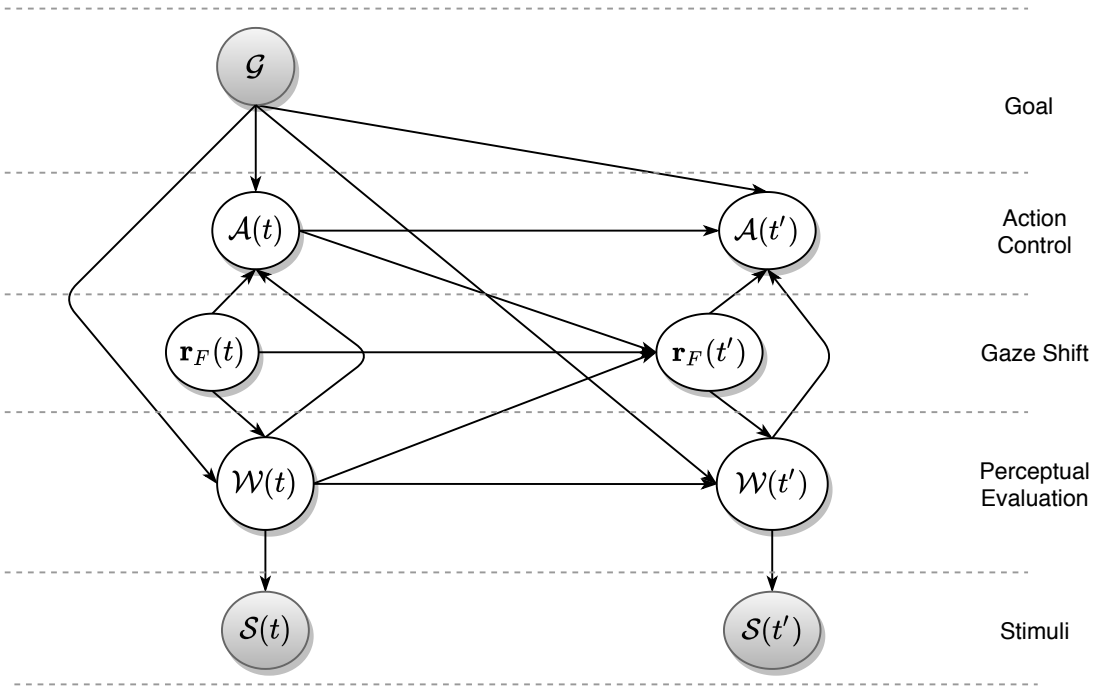


Figure 4.3: *High level view on the adopted model describing the perceptual process accomplished by an ideal perceiver $\mathcal{O}$ when attending a time varying scene. The goal $\mathcal{G}$ of the observer affects both the perceptual evaluation of the stimuli at time t and the action to be taken (where to look next).*

Step 3: Sample where to look next, that is the gaze shift, $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t')$

In a nutshell, the eye guidance loop answers the very question: *Where to look next?* The "where" part (Step 1) concerns the selection of *what* to gaze at - features, objects, actions - and their location within the scene; the "next" part (Steps 2 and 3) involves *how* we gaze at what we have chosen to gaze. The latter crucially brings in the unfolding dynamics of gaze deployment. The overall description of the adopted approach is provided by the Probabilistic Graphical Model depicted in Figure 4.3

As notable, the perceptual process is composed by many layers each of which will be described throughout the present Chapter.

4.2.1 How to define $\mathcal{G}$: the may facets of goals

As a matter of fact, in the real world, most fixations are not generically deployed to objects but allocated to task-relevant objects (Canosa, 2009; Rothkopf et al., 2007; Schütz et al., 2011; Foulsham and Underwood, 2008). The recent theoretical perspectives on active/attentive sensing (Yang et al., 2016) promote the idea that the ultimate objective of the active sensing loop (Steps 1-3) should be to maximize via exploration the long term total rewards and to gain additional knowledge about the environment. Cogently, this endeavour recalls that of animals foraging for food. Animals are likely to choose actions that not only take them closer to known food sources but also yield information about potential new sources (Yang et al., 2016; Averbeck, 2015).

Yet, defining what is a goal is far from evident. The dichotomy between top-down and bottom-up control assumes the former as being determined by the current "endogenous" goals of the observer and the latter as being constrained by the physical, "exogenous" characteristics of the stimuli (independent of the internal state of the observer, e.g., flashes of light, loud noises, etc).

The construct of "endogenous" attentional control (unrelated to stimulus salience) is subtle since it conflates control signals that are "internal" (such as the motivation for paying attention to socially rewarding objects/events), "external" (induced by the given current task voluntarily pursued), and selection history (either learned or evolutionary inherited), which can prioritize items previously attended in a given context.

To discuss thoroughly this point would carry us deep into the study of the complex interaction between cognition and emotion (Pessoa, 2008). A few words must here suffice.

If the ultimate objective of the attentive perceiver is total reward maximisation, one is urged to distinguish between "external" rewards (incentive motivation, e.g, monetary reward) and reward related to "internal" value. Most important for the work presented here, the latter has different psychological facets (Berridge and Robinson, 2003) including affect (implicit "liking" and conscious pleasure) and motivation (implicit incentive salience, "wanting"). Indeed, the selection of socially relevant stimuli by attention has important implications for the survival and wellbeing of an organism, and attentional priority reflects the overall value and the history of such selection (Anderson, 2013).

This also suggests that the crude top-down vs. bottom-up taxonomy of attentional control should be adopted with the uttermost caution (cfr., Awh et al. (2012); Tatler et al. (2011); Groen et al. (2017)).

4.2.2 The neglected perceiver: biases, variability, idiosyncrasy

To date, the vast majority of models have focused on task and stimuli-specific effects but have largely ignored the "observer factor". When considering the *how* component (Steps 2 and 3), though, cogently the perceiver $\mathcal{O}$ is brought in.

On the one hand, regardless of the perceptual input, scan paths exhibit both systematic tendencies and notable inter- and intra-subject variability (c.f.r. Section 2.3.1).

Different individuals move their gaze differently even when confronted with the same task and stimuli. This variability was often treated as noise and data was collapsed across observers. Gaze behavior exhibits individual characteristics much like gait and speech. Unlike gait and speech, gaze behavior strongly determines the visual input arriving in the brain, making such variability an important factor in determining and reflecting one's inner world.

Recent studies examined the variability of eye movements between observers distinguishing which characteristics are stable and reliable, and therefore should be treated as a trait of the observer rather than "noise" (Henderson and Luke, 2014; Bargary et al., 2017). Guy et al. (2019) have reported on a novel gaze related trait that influences the amount of social information accumulated by the observer; it has been shown that the amount of time subjects fixate on others' faces (face-preference) varies between individuals in a reliable manner (Guy et al., 2019)

The variability and bias issues can be explicitly addressed from first principles in the theoretical context of Lévy flights (Brockmann and Geisel, 2000; Boccignone and

Ferraro, 2004). As stated earlier (Section 3.5.3), this direction leeds to treating visual exploration strategies in terms of *foraging* strategies (Wolfe, 2013; Cain et al., 2012; Boccignone and Ferraro, 2014; Clavelli et al., 2014; Napoletano et al., 2015).

Indeed, in certain circumstances, uncertainty may promote almost "blind" visual exploration strategies Tatler and Vincent (2009); Over et al. (2007), much like the behaviour of a foraging animal exploring the environment under incomplete information. As a matter of fact, when animals have limited information about where resource patches are located, different random search strategies can provide different chances to find them Bartumeus and Catalan (2009).

4.2.3 Defining $\mathcal{S}$: the multi-sensory challenge

Humans are multi-sensory perceivers, capable of attentional behaviour on multimodal stimuli, for example those mixing visual and audio stimuli, $\mathcal{S} = \{\mathbf{I}, \mathbf{A}\}$, where $\mathbf{I}$ is a frame sequence and $\mathbf{A}$ an audio signal.

As to computational models, whilst attentional mechanisms have been largely explored for vision systems, there is not much tradition as regards models of attention in the context of sound systems (Kaya and Elhilali, 2017).

The dichotomy between top-down and bottom-up control has been assumed in the auditory attention field of inquiry.

Since the seminal work by Kayser et al. (2005), efforts have been spent to model stimulus-driven attention to the auditory domain, by computing a visual saliency map of the spectrogram of an auditory stimulus (see Kaya and Elhilali (2017) for a comprehensive review).

In this perspective, the combination of both visual and auditory saliencies supporting a multimodal saliency map that grounds multimodal attention becomes a viable route (Onat et al., 2007; Evangelopoulos et al., 2008).

Seminal work on multimodal saliency has been done by Coutrot and Guyader (2014a, 2015, 2016), where static and dynamic low-level visual features were combined with semi-automatically segmented object-based cues (such as faces and annotation of body parts). For the audio track of video frames a speaker diarization technique was proposed based on voice activity detection, audio speaker clustering, and motion detection. This information was then combined with visual information to obtain a saliency map. Clearly, the need for manual face and body part segmentation limits the applicability of these models in real-world scenarios.

In a recent work, Tavakoli et al. (2019) directly learn the mapping using a deep neural network instead of relying on a sampling scheme and multiple feature maps. Their model is distinct from aforementioned audio-visual saliency models because applicable to any scene type, not only to conversational videos, and it is a single end-to-end trainable framework for the multi-modal saliency prediction.

4.3 Overview of the basic model architecture

The general problem addressed by the proposed model may be stated as follows:

The dynamic multimodal landscape $\mathcal{W}(t)$, the world as perceived by subject $\mathcal{O}$, is a "patchy" environment. Patches are clumps of audio-visual information to which gaze is deployed. The perceiver scrutinises "items" within a patch and, at any time t, makes

action decisions $\mathcal{A}(t)$ as to: 1) which patches are to be inspected; 2) when to leave the patch currently visited for focussing on a new patch. In this endeavour, the unfolding of gaze deployment, $\mathbf{r}(t) \rightarrow \mathbf{r}(t')$, alternates between scanning the patch, for probing and exploiting the chunks of information locally available, and longer, explorative relocations between patches.

To frame such problem, in essence a foraging problem, we make a number of assumptions.

A1 The unfolding of gaze deployment in time, is best described as a stochastic process, namely a biased random walk of a forager over the changing landscape (cfr. Fig. 4.1)

The landscape $\mathcal{W}(t)$ generated by $\mathcal{O}$ from the audio-visual stream $\mathcal{S}(t) = \{\mathbf{I}(t), \mathbf{A}(t)\}$ is inherently stochastic and the observer has partial information, since patches may change unpredictably in time. Further, as discussed in Section 4.2, we need to take into account $\mathcal{O}$'s variability and biases.

A2 The gaze walk can be accounted for by one and only model of oculomotor behavior, namely an Ornstein-Uhlenbeck process; the process acts at different scales, from landscape exploration to local patch exploitation.

Indeed, recent work has been challenging the view that exploration and fixation are dichotomous. Current literature suggests instead that visual fixation is functionally equivalent to visual exploration on a spatially focused scale (Otero-Millan et al., 2013; Martinez-Conde et al., 2013). In brief, they are two extremes of a functional continuum. Interestingly enough, recent experiments confirmed scale invariance in the temporal structure of the larger shifts in gaze position (saccades), which has also been observed in fixational eye movements while the eye is gauging a localized region in the visual field (Marlow et al., 2015). A consequence of this assumption is that we do not need to rely, in our analysis, on gaze behaviour classification such as saccade, fixation, and smooth pursuit.

A3 In a multimodal landscape conveying social content, the forager's random walk for exploration/exploitation is modulated by the value $\mathbf{v}$, which is internally assigned by the agent $\mathcal{O}$ to socially rewarding items. Value dynamics can thus be inferred from the oculomotor behaviour of real subjects.

In the work presented here, no "external" task is assigned to the perceiver; thus, value $\mathbf{v}$ is modulated by the "internal" drive towards spotting socially relevant objects/events.

In a landscape featuring social content, the most prominent visual objects are likely to be faces and audio objects as represented by speakers' voices (Foulsham et al., 2010). These, eventually, will maximally contribute to the relevant patches within $\mathcal{W}(t)$ that will bias the random walk of the perceiver's gaze.

Under such circumstances, gaze deployment is obtained as follows. Along a *pre-attentive stage*, audio-visual features are derived to assess the likelihood of the spatio-temporal occurence of such events. This provides the basis for setting up time-varying priority maps $\mathbf{L}_\ell$ ($\ell = 1, \cdots, N_\ell$) and for gauging their moment-to-moment value v_ℓ

in the context of the scene. From priority maps, a number of value-based patches $\mathcal{P}_p^{(\ell)}$ ($p = 1, \cdots, N_P$) are generated.

The *attentive stage* is distilled in the evolution of the gaze state represented by point $\mathbf{r}_F(t) = (x_F(t), y_F(t))^T$ in a continuous 2-dimensional space, at any time $t \geqslant 0$, which sets the focus of attention (FoA). As such, gaze dynamics $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t')$ unfolding in time defines a trajectory, which is the realisation $R_F(t) = \mathbf{r}(t)$, of a continuous-time stochastic process $\{R_F(t) : t \geqslant 0\}$. From now on, for sake of simplicity and with some abuse of notation, we shall use $\mathbf{r}_F(\cdot)$ for denoting both the process/random variable and its realization; the same holds for other random variables, unless otherwise specified.

The process is conceived as an OU process operating at two different scales. These parametrise local and global biased random walks so that area-concentrated phases within patches (exploitation) alternate with large distance relocation phases between patches (exploration).

The switch between the two states of oculomotor behaviour, patch exploitation and landscape exploration, is provided by a foraging decision resulting from comparing the expected reward gained within currently exploited patch against the average reward that could be gained moving to other patches available within the landscape. If exploration is undertaken, then the choice of a new patch must be made. State switching and patch choice are the behavioural decisions $\mathcal{A}$ available to the forager.

A further assumption of the model presented here relates to the patch exploitation mechanism. In stochastic foraging theory, the time spent within a patch depends on the potential value of a patch, which is based on the the expected rate, the forager's current expectations on the number of items in the patch and how easy they should be to find, (Green, 1980; McNamara, 1982; McNamara and Houston, 1985). In the case of internal goals, it is difficult to exactly define what is an item. For example, consider a patch embedding a speaker's face. Items could either be main facial shape features (eyes, nose, etc.), or action units of facial expressions, or joint lip movements and spoken words, etc. Even if we could count the items, we would not know how many items are processed when gaze is deployed at point $\mathbf{r}(t)$ in the course of local patch exploration; multiple items might be processed in parallel (Ehinger and Wolfe, 2016).

On this basis, in the same vein of the foraging literature and its applications in perception (Wolfe, 2013; Cain et al., 2012; Ehinger and Wolfe, 2016), our model abstracts from the actual mechanisms of specific gaze behaviour within a region of interest under a given task, but isolates some very relevant phenomenological aspects akin to be shaped in statistical terms. This suits our needs, our concern here being with the general view rather than with the details.

Patches and items within the patch are encountered according to a Poisson process; indeed Poisson processes and associated exponential waiting times play an important role to relate points of gaze and global/local scene characteristics (Barthelmé et al., 2013; Han et al., 2013). Here, patches are modelled as independent Poisson process generators. Number of items are sampled from a Poisson distribution, which allows to derive a simple law for estimating the instantaneous information gain of the perceiver within the patch and to compare the latter with the average gain achievable over the landscape. This provides a sound basis for deciding when to relocate to another patch and how to choose the next patch to be exploited, namely the actions $\mathcal{A}(t)$ moment-by-

moment available to the perceiver.

The overall control algorithm for gaze deployment is summarised in the *GazeDeploy* procedure outlined in Algorithm 1. Its steps are detailed in the following sections and a Python implementation of the procedure is freely available on GitHub[1]

However, at this point, Figure 4.4 might provide a useful insight of the overall behaviour of the procedure.

Given an input conversational clip, summarised as an excerpt of four subsequent frames (top to bottom, left column), the *GazeDeploy* procedure outputs a continuous gaze trajectory as generated by one artificial observer (second column), whilst the third column shows the focus of attention (FoA) set at the corresponding time. In the top, the second and the bottom raw the simulated observer scrutinises the current speaker, as espected, whilst in the third and fourth raws, a brief glance is deployed to the woman listening on the left and to the onset of the hand gesture of the forthcoming speaker.

It is worth remarking that one such individual trajectory might stochastically deviate to some extent from those of other observers, either real or artificial. This can be appreciated from the fourth and the last columns. These represent the time-varying fixation maps computed from a paired number of either artificial observers and actual human observers. Note that when the conversational scene becomes more complex (typically due to people arguing, gesturing, turn-taking, etc.), the fixation maps are characterised by higher spatio-temporal dispersion, which is a signature of the attention variability of observers. Such uncertainty is captured by both the artificial and actual maps. In such circumstances, indeed, the inter-observer variability grows, and individual observers are likely to be driven their own expectation and other idiosyncratic factors.

4.4 The preattentive stage: perceiving the audio-visual landscape and it's value

At the heart of the time-varying, pre-attentive perceptual representation $\mathcal{W}(t)$ lies the concept of priority map. Intuitively, a priority map $\mathbf{L}$ combines top-down (relevance under given goals $\mathcal{G}$) and bottom-up (salience) mechanisms for eye guidance (Desimone and Duncan, 1995; Egeth and Yantis, 1997; Serences and Yantis, 2006; Fecteau and Munoz, 2006). More generally, it can be conceived as a dynamic map of the perceptual landscape constructed from a combination of properties of the external stimuli, intrinsic expectations, and contextual knowledge (Chikkerur et al., 2010; Torralba, 2003); it can also be designed to act as a form of short term memory to keep track of which potential targets have been attended. As such, the representation entailed by a priority map differs from that provided at a lower level by feature maps $\mathbf{X}$ (or classic salience).

Priority maps are used in our model to sample the audio-visual patches of interest that define the perceiver's landscape. Each patch bears a value inherited from its priority map. Here, rather than shaping value in the form of a map (in a sense, a further instance of a priority map, see Klink et al. (2014); Chelazzi et al. (2014)), we consider it as a process that moment to moment weighs the relevance of the the different priority maps conditionally on the observer's goal. Under such circumstances, we generally assume attention as driven by goals $\mathcal{G}$ that, in turn, set the appropriate value $\mathcal{V}$ to events/objects occurring in the audiovisual scene. Also, in the work presented here, we assume that no

[1] `https://github.com/phuselab/GazeDeploy`

Algorithm 1 Gaze control in a multimodal landscape

Input: Visual stream $\{\mathbf{I}\}$, audio stream $\{\mathbf{A}\}$, goals $\mathcal{G}$ (internal or external), T the duration to be simulated, video fps FPS, random walk sampling rate fs.

Output: Prediction of gaze

procedure GAZEDEPLOY
$\delta t = \frac{1}{FPS}, \quad \delta u = \frac{1}{fs}$
Initialisation of first gaze location $\mathbf{r}(t_1)$ on patch $p = k$, with behavioural state $s(t_1) = 1$ {*Exploitation mode*}
for $n = 2$ **to** $\frac{T}{\delta t}$ **do**
  {**Preattentive feedforward stage**}
  Compute the current state of the perceptual landscape, in terms of audio-visual priority maps $\{\mathbf{L}_\ell\}$ and distributions $\{\mathcal{L}_\ell(t_n)\}$ (Eqs. 4.1, 4.2)
  {**Value inference**}
  Infer value dynamics $\{v_\ell(t_n)\}$ given all available information $\mathcal{I}(t_1 : t_n)$ up to time t_n, (Eq. 4.3)
  {**Landscape evaluation**}
  Compute audio-visual patches $\{\mathcal{P}_p(t_n)\}$ as potential value-sensitive attractors
  Compute the expected average gain $Q(t_n)$ from all patches in the landscape (Eq. 4.19)
  {**Attentive stage**}
  if $s(t_n) = 1$ **then**
    {**Exploitation: patch handling**}
    Set the parameters $\boldsymbol{\mu}_p^{(s_t)}, \boldsymbol{\Psi}_p^{(s_t)}$ for OU sampling according to state $s(t_n)$ and current patch indexed by $p(t_n)$
    while within patch **do**
      {**Exploitation: local gaze shifting** }
      for $j = 0$ **to** $\frac{fs}{FPS}$ **do**
        Sample the OU gaze relocation $\mathbf{r}_{t_{n-1}+(j\times\delta u)} \rightarrow \mathbf{r}_{t_{n-1}+(j+1\times\delta u)}$
      end for
      {**Behavioural state sampling**}
      Compute the instantaneous expected gain $g_p(t_{W_p})$ for current patch (Eq. 4.17)
      Compare current patch gain against the expected average gain Q from the environment (Eq. 4.20)
      Sample the behavioural state $s(t_n)$ at time $t_n = t_{n-1} + \delta t$ (Eq. 4.11)
    end while
  else
    {**Exploration: patch-choice**}
    Sample next most valuable attractor $p(t_{n-1} + \delta t)$ (Eq. 4.12)
    Set the parameters $\boldsymbol{\mu}_p^{(s_t)}, \boldsymbol{\Psi}_p^{(s_t)}$ for OU sampling according to state $s(t_n)$ and attractor $p(t_n)$
    {**Exploration: relocation gaze shifting** }
    for $j = 0$ **to** $\frac{fs}{FPS}$ **do**
      Sample the OU gaze relocation $\mathbf{r}_{t_{n-1}+(j\times\delta u)} \rightarrow \mathbf{r}_{t_{n-1}+(j+1\times\delta u)}$
    end for
  end if
end for
end procedure=0

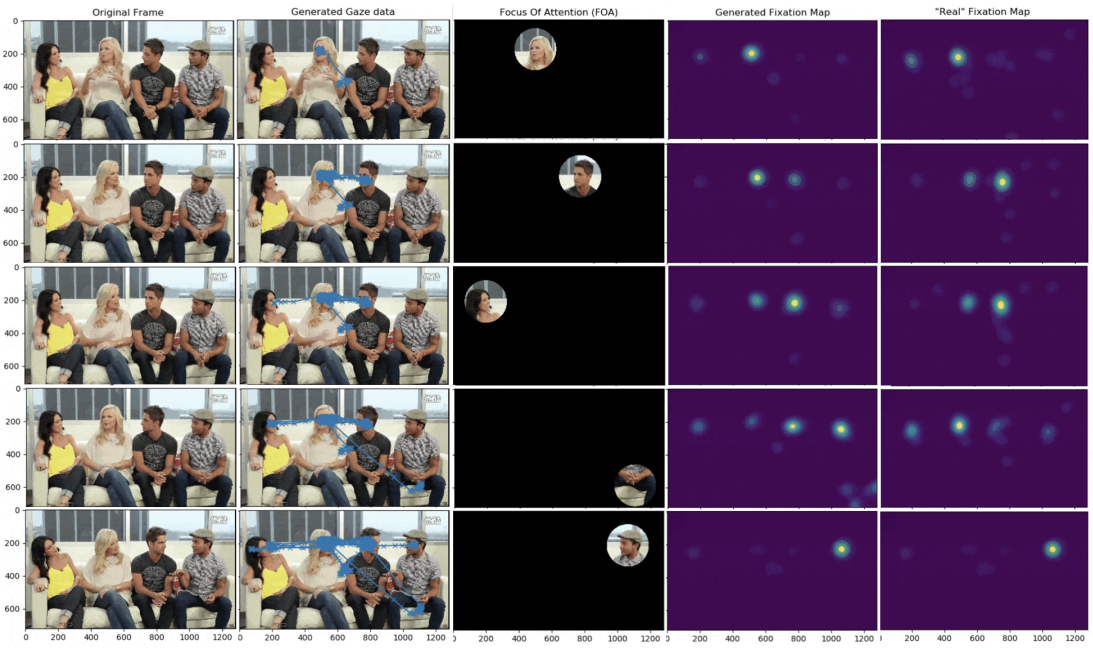


Figure 4.4: *The behaviour of the* GazeDeploy *procedure captured through the excerpt of four subsequent frames of a conversational clip. The left-most column summarises the input sequence (top to bottom). The second column displays the output of the procedure, namely the continuous gaze trajectory (graphically overlapped on the input frame) as generated by one artificial observer up to that frame. The third column highlights the focus of attention (FoA) set on the scene. To weigh such individual trajectory in the context of other observers' behaviour, the fourth and right-most columns represent the time-varying fixation maps (a.k.a, heatmaps, attentional maps) computed from a paired number of either artificial observers and actual human observers, respectively.*

explicit task is assigned to the perceiver; thus, value $\mathcal{V}$ is modulated by the "internal" goal (drive) towards spotting socially relevant objects/events.

4.4.1 Computing priority maps

Perceptual spatial attention driven by multimodal cues mainly relies on visual and audio-visual priority maps, which we define as the RVs $\mathbf{L}_V$ and $\mathbf{L}_{AV}$, respectively. Formally, a priority map $\mathbf{L}$ is the matrix of binary RVs $l(\mathbf{r})$ denoting if location $\mathbf{r}$ is to be considered relevant ($l(\mathbf{r}) = 1$) or not ($l(\mathbf{r}) = 0$), with respect to possible visual or audio-visual "objects" occurring within the scene. Thus, given the video and audio streams defining the audio-video landscape, $\{\mathbf{I}(t)\}$, $\{\mathbf{A}(t)\}$, respectively, a preliminary step is to evaluate, given two time instants $t < t'$, the posterior distributions $P(\mathbf{L}_V(t') \mid \mathbf{L}_V(t), \mathbf{I}(t'))$ and $P(\mathbf{L}_{AV}(t') \mid \mathbf{L}_{AV}(t), \mathbf{A}(t'), \mathbf{I}(t'))$, where $t' - t = \delta t$ with δt being an arbitrary time step.

The steps behind such estimate can be derived by resorting to the conditional dependencies defined in the PGM in Fig. 4.5.

Backward inference $\{\mathbf{A}(t), \mathbf{I}(t)\} \rightarrow \{\mathbf{L}_V(t), \mathbf{L}_{AV}(t)\}$ stands upon a set of perceptual features $\mathbf{F}(t) = \{f(t)\}$ that can be estimated from the multimodal stream. From now on, for notational simplicity, we will omit time indexing t, unless needed.

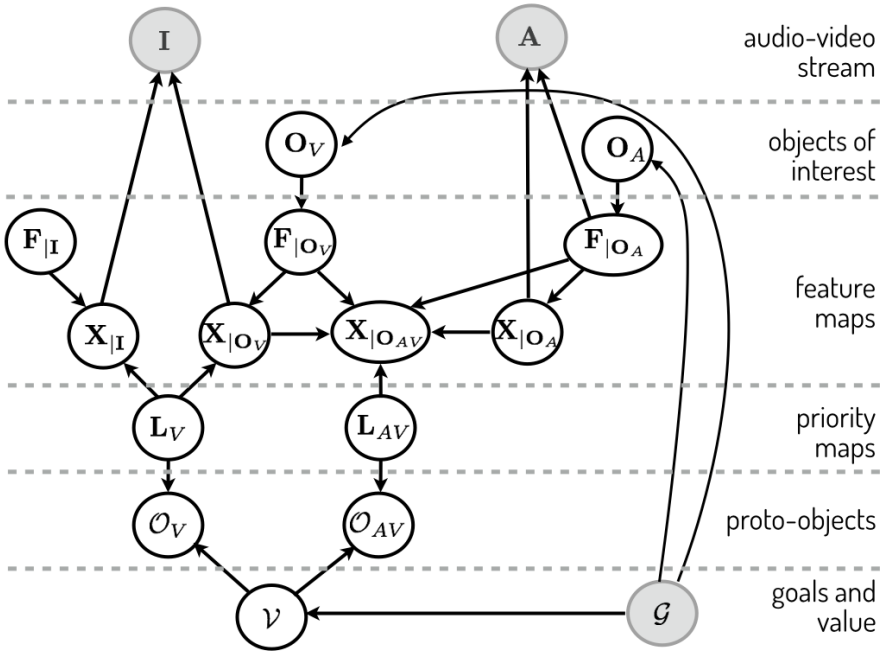


Figure 4.5: *An overall view of the model as a Probabilistic Graphical Model describing the computation of the audio-visual priority maps. This can be seen as a zoom-in on the* ***Perceptual Evaluation*** *layer appearing in the PGM of Figure 4.3. Time index t has been omitted for simplicity.*

As to the visual stream, we distinguish between two kinds of visual features: generic features $\mathbf{F}_{|\mathbf{I}}$ - such as edge, texture, colour, motion features-, and object-dependent features, $\mathbf{F}_{|\mathbf{O}}$. As to object-based features, these are to be learned by specifically taking into account the classes of objects that are likely to be relevant under the goal $\mathcal{G}$, via the distribution $P(\mathbf{O} \mid \mathcal{G})$. Here, where the task is free viewing/listening, and internal goals are biased towards social cues, the prominent visual objects are faces, $\mathbf{O}_V = \{face\}$. Both kinds of visual features, $\mathbf{F}_{|\mathbf{I}}$and $\mathbf{F}_{|\mathbf{O}}$, can be estimated in a feed-forward way. Note that in the literature face information is usually referred to as a top-down cue (Schütz et al., 2011) as opposed to bottom-up cues. However, much like physically driven features, they are phyletic features, and their distribution $P(\mathbf{F}_{|\mathbf{O}_V} \mid \mathbf{O}_V = face)$ is learnt by biological visual systems along evolution or in early development stages (Wilkinson et al., 2014).

In order to be processed, features $\mathbf{F}_{|\mathbf{I}}$ and $\mathbf{F}_{|\mathbf{O}_V}$ need to be spatially organised in feature maps. A feature map $\mathbf{X}$ is a topographically organised map that encodes the joint occurrence of a specific feature at a spatial location (Chikkerur et al., 2010). It can be considered the probabilistic counterpart of a salience map (Chikkerur et al., 2010) and it can be equivalently represented as a unique map encoding the presence of different object dependent features $\mathbf{F}_{f|\mathbf{O}_V}$, or a set of object-specific feature maps, i.e. $\mathbf{X} = \{\mathbf{X}_f\}$ (e.g., a face map, a body map, etc.). More precisely, $\mathbf{X}_f$ is a matrix of binary RVs $x(\mathbf{r})$ denoting whether feature f is present or not present at location $\mathbf{L} = \mathbf{r}$. Simply put, $\mathbf{X}_f$ is a map defining the spatial occurrence of $\mathbf{F}_{f|\mathbf{O}_V}$ or $\mathbf{F}_{f|\mathbf{I}}$. In our case, we need to estimate the posteriors $P(\mathbf{X}_{|\mathbf{I}} \mid \mathbf{F}_{|\mathbf{I}})$ and $P(\mathbf{X}_{|\mathbf{O}_V} \mid \mathbf{F}_{|\mathbf{O}_V})$.

As to the processing of audio, similarly to visual processing, auditory objects form across different analysis scales (Shinn-Cunningham, 2008). Formation of sound elements with contiguous spectro-temporal structure, relies primarily on local structures (e.g., onsets and offsets, harmonic structure, continuity of frequency over time), while social communication signals, such as speech, have a rich spectro-temporal structure supporting short-term object formation (e.g. formation of syllables). The latter are linked together over time through continuity and similarity of higher-order perceptual features, such as location, pitch, timbre and learned meaning.

In our setting, the objects of interest $\mathbf{O}_A$ are represented by speakers' voices (Foulsham et al., 2010), and features $\mathbf{F}_{f|\mathbf{O}_A}$ suitable to represent speech cues. In this work, we are not considering other audio sources (e.g, music). From a social perspective, we are interested in inferring the audio-visual topographic maps of speaker/non-speakers, $\mathbf{X}_{|\mathbf{O}_{AV}}$, given the available faces in the scene and speech features via the posterior distribution $P(\mathbf{X}_{|\mathbf{O}_{AV}} \mid \mathbf{X}_{|\mathbf{O}_A}, \mathbf{X}_{|\mathbf{O}_V}, \mathbf{F}_{|\mathbf{O}_A}, \mathbf{F}_{|\mathbf{O}_V})$, where $\mathbf{X}_{|\mathbf{O}_{AV}} = x(\mathbf{r})$ denotes whether a speaker/non-speaker is present or not present at location $\mathbf{r}$.

At this point, audio-visual perception has been cast in a spatial attention problem and priority maps $\mathbf{L}_V$and $\mathbf{L}_{AV}$ can be eventually estimated through distributions $P(\mathbf{L}_V(t') \mid \mathbf{L}_V(t), \mathbf{X}_{|\mathbf{I}}, \mathbf{X}_{|\mathbf{O}_V})$ and $P(\mathbf{L}_{AV}(t')|\mathbf{L}_{AV}(t), \mathbf{X}_{|\mathbf{O}_{AV}})$.

Note that, in general, the representation entailed by a priority map differs from that provided at a lower level by feature maps $\mathbf{X}$ (or classic salience). It can be conceived as a dynamic map of the perceptual landscape constructed from a combination of properties of the external stimuli, intrinsic expectations, and contextual knowledge (Chikkerur et al., 2010; Torralba, 2003). Also, it can be designed to act as a form of short term memory to keep track of which potential targets have been attended. Thus, $\mathbf{L}(t')$ depends on both current perceptual inferences on feature maps at time t' and priority at time $t < t'$. Denote:

$$\begin{aligned}\pi_{AV} &= P(\mathbf{X}_{|\mathbf{O}_{AV}} \mid \mathbf{X}_{|\mathbf{O}_A}, \mathbf{X}_{|\mathbf{O}_V}, \mathbf{F}_{|\mathbf{O}_A}, \mathbf{F}_{|\mathbf{O}_V}) \\ \pi_V &= P(\mathbf{X}_{|\mathbf{O}_V} \mid \mathbf{F}_{|\mathbf{O}_V})\end{aligned}$$

the distributions related to the feature maps, and:

$$\begin{aligned}\mathcal{L}_V(t') &= P(\mathbf{L}_V(t') \mid \mathbf{L}_V(t), \mathbf{X}_{|\mathbf{I}}, \mathbf{X}_{|\mathbf{O}_V}) \\ \mathcal{L}_{AV}(t') &= P(\mathbf{L}_{AV}(t')|\mathbf{L}_{AV}(t), \mathbf{X}_{|\mathbf{O}_{AV}})\end{aligned}$$

the distributions related to the priority maps. Then, the latter can be estimated as:

$$\mathcal{L}_V(t') = \alpha_V \pi_V(t') + (1 - \alpha_V)\mathcal{L}_V(t), \tag{4.1}$$

$$\mathcal{L}_{AV}(t') = \alpha_{AV} \pi_{AV}(t') + (1 - \alpha_{AV})\mathcal{L}_{AV}(t). \tag{4.2}$$

where α_V and α_{AV} weight the contribution of currently estimated feature maps with respect to previous priority maps.

Priority map dynamics requires an initial prior $P(\mathbf{L})$, which can be designed to account for spatial tendencies in the perceptual process; for instance, human eye-tracking

studies have shown that gaze fixations in free viewing of dynamic natural scenes are biased toward the center of the scene ("center bias"), (Tatler and Vincent, 2009; Le Meur and Coutrot, 2016), which can be modelled by assuming a Gaussian distribution located on the viewing center.

4.4.2 Deriving Feature Maps

The input stimuli $\mathcal{S}$ are represented by the time-varying visual and audio streams, $\mathcal{S}(t) = \{\mathbf{I}(t), \mathbf{A}(t)\}, t = 1, \cdots, T$, where $\mathbf{I}$ is the frame sequence and $\mathbf{A}$ the audio signal.

In order to derive a priority map, we need to specify which features $\mathbf{F}$ are to be taken into account, given the context or goal $\mathcal{G}$, and the feature maps $\mathbf{X}$, that is the topographically organised maps that encode the joint occurrence of a specific feature at a spatial location (Chikkerur et al., 2010). In a probabilistic setting, a feature map $\mathbf{X}_f$ is a matrix of binary RVs $x(\mathbf{r})$ denoting whether feature f is present or not present at location $\mathbf{L} = \mathbf{r}$ (Chikkerur et al., 2010). It can be equivalently represented as a unique map encoding the presence of different object dependent features $\mathbf{F}_{f,\mathbf{O}}$, or a set of object-specific feature maps, i.e. $\mathbf{X} = \{\mathbf{X}_f\}$ (e.g., in the visual realm, a face map, a body map, etc.)

Visual features

From input $\mathbf{I}$, two kinds of visual features are derived: generic visual features $\mathbf{F_I}$ - such as edge, texture, colour, motion features-, and object-dependent features, $\mathbf{F}_{\mathbf{O}_V}$. The latter are selected by taking into account the classes of objects that are likely to be relevant under the goal $\mathcal{G}$.

Internal goals are biased towards social cues, thus the prominent visual objects are faces, $\mathbf{O}_V = \{face\}$. Both kinds of visual features, $\mathbf{F_I}$ and $\mathbf{F}_{\mathbf{O}_V}$, can be estimated in a feed-forward way.

In order to be processed, features $\mathbf{F_I}$ and $\mathbf{F}_{\mathbf{O}_V}$ need to be spatially organized in feature maps. In the visual attention context, the distribution $P(\mathbf{X})$ can be considered the probabilistic counterpart of the classic salience map (Chikkerur et al., 2010). Thus, $\mathbf{X}_{f,\mathbf{I}}$ represents the support of a low-level saliency map, whilst $\mathbf{X}_{f,\mathbf{O}_V}$ is the support of an high-level, object-based saliency map.

At this stage, the inferential step entails estimating the posteriors $P(\mathbf{X_I} \mid \mathbf{F_I})$ and $P(\mathbf{X}_{\mathbf{O}_V} \mid \mathbf{F}_{\mathbf{O}_V})$, whatever the technique adopted.

In order to derive the physical stimulus feature map $\mathbf{X_I}$, we rely on the spatio-temporal saliency method proposed by Seo and Milanfar (2009) based on local regression kernel center/surround features. It avoids specific optical flow processing for motion detection and has the advantage of being insensitive to possible camera motion. By assuming uniform prior on all locations, the evidence from a location $\mathbf{r}$ of the frame is computed via the likelihood $P(\mathbf{I}(t) \mid \mathbf{x}_f(\mathbf{r}, t) = 1, \mathbf{F_I}, \mathbf{r}_F(t)) = \frac{1}{\sum_s} \exp\left(\frac{1-\rho(\mathbf{F}_{\mathbf{r},c}, \mathbf{F}_{\mathbf{r},s})}{\sigma^2}\right)$, where $\rho(\cdot) \in [-1, 1]$ is the matrix cosine similarity (see Seo and Milanfar (2009), for details) between center and surround feature matrices $\mathbf{F}_{\mathbf{r},c}$ and $\mathbf{F}_{\mathbf{r},s}$ computed at location $\mathbf{r}$ of frame $\mathbf{I}(t)$.

The visual object-based feature map $\mathbf{X}_{\mathbf{O}_V}$ entails a face detection step. There is a huge number of methods currently available: the one proposed Hu and Ramanan (2017)

has shown, in our preliminary experiments, to bear the highest performance. It relies on a feed-forward deep network architecture for scale invariant detection. Starting with an input frame $\mathbf{I}(t)$, a coarse image pyramid (including interpolation) is created. Then, the scaled input is fed into a Convolutional Neural Network (CNN) to predict template responses at every resolution. Non-maximum suppression (NMS) is applied at the original resolution to get the final detection results. Their confidence value is used to assign the probability $P(\mathbf{X}_{\mathbf{O}_V} \mid \mathbf{F}_{\mathbf{O}_V}, \mathbf{L}_V = \mathbf{r})$ of spotting face features $\mathbf{F}_{\mathbf{O}_V}$ at $\mathbf{L}_V = \mathbf{r}$, according to a gaussian distribution located on the face center modulated by detection confidence and face size.

Audio and audio-visual features

From input $\mathbf{A}$, auditory objects form across different analysis scales (Shinn-Cunningham, 2008). Formation of sound elements with contiguous spectro-temporal structure, relies primarily on local structures (e.g., onsets and offsets, harmonic structure, continuity of frequency over time), while social communication signals, such as speech, have a rich spectro-temporal structure supporting short-term object formation (e.g. formation of syllables). The latter are linked together over time through continuity and similarity of higher-order perceptual features, such as location, pitch, timbre and learned meaning.

In our setting, the objects of interest $\mathbf{O}_A$ are represented by speakers' voices (Foulsham et al., 2010), and features $\mathbf{F}_{f,\mathbf{O}_A}$ suitable to represent speech cues. In this work, we are not considering other audio sources (e.g, music).

From a social perspective, we are interested in inferring the audio-visual topographic maps of speaker/non-speakers, $\mathbf{X}_{\mathbf{O}_{AV}}$, given the available faces in the scene and speech features via the posterior distribution $P(\mathbf{X}_{\mathbf{O}_{AV}} \mid \mathbf{X}_{\mathbf{O}_A}, \mathbf{X}_{\mathbf{O}_V}, \mathbf{F}_{\mathbf{O}_A}, \mathbf{F}_{\mathbf{O}_V})$,

where $\mathbf{X}_{|\mathbf{O}_{AV}} = x(\mathbf{r})$ denotes whether a speaker/non-speaker is present or not present at location $\mathbf{r}$.

Technically, the features $\mathbf{F}_{|\mathbf{O}_A}$ used to encode the speech stream are the Mel-frequency cepstral coefficients (MFCC). The Mel-frequency cepstrum is highly effective in speech recognition and in modelling the subjective pitch and frequency content.

The audio feature map $\mathbf{X}_{\mathbf{O}_A}(t)$ can be conceived as a spectro-temporal structure computed from a suitable time window of the audio stream, representing MFCC values for each time step and each Mel frequency band. It is important to note, that the problem of deriving the speaker/non-speaker map $\mathbf{X}_{\mathbf{O}_{AV}}$ when multiple faces are present, is closely related to the AV synchronisation problem (Chung and Zisserman, 2016); namely, that of inferring the correspondence between the video and the speech streams, captured by the joint probability $P(\mathbf{X}_{\mathbf{O}_{AV}}, \mathbf{X}_{\mathbf{O}_A}, \mathbf{X}_{\mathbf{O}_V}, \mathbf{F}_{\mathbf{O}_A}, \mathbf{F}_{\mathbf{O}_V}, \mathbf{L}_{AV})$.

The speaker's face is the one with the highest correlation between the audio and the video feature streams, whilst a non-speaker should have a correlation close to zero. It has been shown that the synchronisation method presented in Chung and Zisserman (2016) can be extended to locate the speaker vs. non-speakers and to provide a suitable confidence value. The method relies on a two-stream CNN architecture (SynchNet) that enables a joint embedding between the sound and the face images. In particular we use the Multi-View version (Chung and Zisserman, 2016, 2017)), which allows the speaker identification on profile faces and does not require explicit lip detection. To such end, 13 Mel frequency bands are used at each time step, where features $\mathbf{F}_{\mathbf{O}_A}(t)$ are computed at sampling rate for a 0.2-secs time-window of the input signal $\mathbf{A}(t)$. The

same time-window is used for the video stream input.

4.4.3 Inferring the value of preattentive information

Attentional value is set by the "internal" goal (drive) $\mathcal{G}$ towards spotting socially relevant objects/events occurring in the scene. As such, it is a hidden state of the perceiver. The problem we are facing now is to set up an inferential procedure so that, given all available information from the onset of the process up to time t, say $\mathcal{I}(1:t)$, the latent value $\mathbf{v}(t)$ can be estimated,

$$\mathbf{v}(t) \mid \mathcal{I}(1:t) \sim P(\mathcal{I}(1:t)). \tag{4.3}$$

Information $\mathcal{I}(t)$ should encompass both perceivers' behaviour and stimulus content. Consider that, on the one hand, we know that the actual moment-to-moment deployment of attention over the landscape is the outcome of a value assignment procedure. We assume that the result of attention allocation is summarised through the time-varying heatmap $\mathcal{H}(t)$, which can be easily computed from eye-tracked gaze positions (fixations) of the perceivers Bylinskii et al. (2019). On the other hand, the information available from the stimulus is, at this point, pre-attentively captured via densities $\mathcal{L}_\ell(t)$. Recall that a priority map density $\mathcal{L}_\ell(t)$ can be conceived as a dynamic predictor of potential gaze allocation in space. We surmise that each map contributes to such prediction conditionally on the value it bears for the observer at moment t.

Formally, define $\mathbf{v}(t) = (v_1(t) \cdots v_{N_\ell}(t))^T$ the time-varying random vector of values that are internally assigned to priority map densities $\mathcal{L}_\ell(t)$. Under such circumstances, the mapping $\mathcal{H}(t) = h(\{\mathcal{L}_\ell(t)\}, \mathbf{v}(t))$ can be simply cast in terms of the linear regression equation

$$\mathcal{H}(t) = \sum_\ell v_\ell(t)\mathcal{L}_\ell(t) + \omega(t), \tag{4.4}$$

which specifies the observers' heatmap $\mathcal{H}(t)$ as the linear combination of predictors (regressors) derived from the stimulus, namely the priority densities $\mathcal{L}_\ell(t)$, perturbed by noise $\omega(t)$. Here, $\mathcal{H}(t)$ is a $2D$ matrix having dimensions equal to the dimensions of the $\mathcal{L}_\ell(t)$ matrices. Eq. 4.4 specifies a time-varying linear regression, since $v_\ell(t)$ are unknown time-varying coefficients. A straightforward dynamics for the latter is to let $v_\ell(t)$ vary over time according to a random walk, where the value displacement $dv_\ell(t)$ simply amounts to a Brownian displacement $dW_\ell(t)$, i.e. $dv_\ell(t) = dW_\ell(t)$.

Then, the dynamic regression model can be conveniently written in terms of the following vector state-space model:

$$\mathbf{h}(t) = \mathbf{P}(t)\mathbf{v}(t) + \boldsymbol{\omega}(t), \quad \boldsymbol{\omega}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}(t)) \tag{4.5}$$

$$\mathbf{v}(t) = \mathbf{v}(t - \delta t) + \boldsymbol{\epsilon}(t), \quad \boldsymbol{\epsilon}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}(t)) \tag{4.6}$$

where: $\mathbf{h}(t) = vec(\mathcal{H}(t))$ is the observation vector of dimension $|\mathcal{H}| \times 1$, obtained by vectorising matrix $\mathcal{H}$; $\mathbf{P}(t) = [vec(\mathcal{L}_1(t)) \mid \cdots \mid vec(\mathcal{L}_{N_l}(t))]$ is the matrix whose columns are the vectorised predictors.

The Gaussian disturbances, namely, the process noise $\boldsymbol{\epsilon}(t)$ (with $\mathbf{Q} = cov(\mathbf{v})$) and the observation noise $\boldsymbol{\omega}(t)$ (with $\mathbf{R} = cov(\mathbf{h})$) are both serially independent and also independent of each other.

Online inference of value (Eq. 4.3) can eventually be performed by solving the filtering problem $P(\mathbf{v}(t) \mid \mathbf{h}(1:t))$ under Markov assumption, where $\mathbf{h}$ is a function of the priority map distributions $\mathcal{L}_\ell$ via the observation/regression in Eq. 4.5. This way, current goal and selection history effects are both taken into account (Awh et al., 2012).

4.4.4 Sampling value sensitive patches

Priority maps and related values are then used for patch sampling. Patches formalise the concept of multimodal attention attractors and inherit the value from the generating priority maps.

Much like proto-objects postulated by object-based attention approaches, they represent the dynamic interface between attentive and pre-attentive processing (Boccignone and Ferraro, 2014).

Given a priority map $\mathbf{L}_\ell$, the spatial support of possible patches is computed.

Denote $\mathcal{M}_p^{(\ell)} = \{m_p^{(\ell)}(\mathbf{r})\}_{\mathbf{r}\in\mathbf{L}^{(\ell)}}$ the map of binary RVs indicating the presence or absence of a patch p.

Assume independent patches, within and across priority maps $\mathbf{L}_\ell$. The map of patches generated by $\mathbf{L}_\ell$ is defined as $\mathcal{M}^{(\ell)} = \bigcup_{p=1}^{N_P^{(\ell)}} \mathcal{M}_p^{(\ell)}$, where $\mathcal{M}_p^{(\ell)} \bigcap \mathcal{M}_k^{(\ell)} = \emptyset, p \neq k$ and the overall patch support map is $\mathcal{M} = \bigcup_{\ell=1}^{N_{(\ell)}} \mathcal{M}^{(\ell)}$.

To derive patches from priority maps we need first to estimate their support $\mathcal{M}(t) = \{m(\mathbf{r},t)\}_{\mathbf{r}\in L}$, such that $m(\mathbf{r},t) = 1$ if $\mathcal{L}_\ell(t) > T_M$, and $m(\mathbf{r},t) = 0$ otherwise. The threshold T_M is adaptively set so as to achieve 90% significance level in deciding whether the given priority values are in the extreme tails of the pdf $\mathcal{L}_\ell$. The procedure is based on the assumption that an informative patch is a relatively rare region and thus results in values which are in the tails of the distribution.

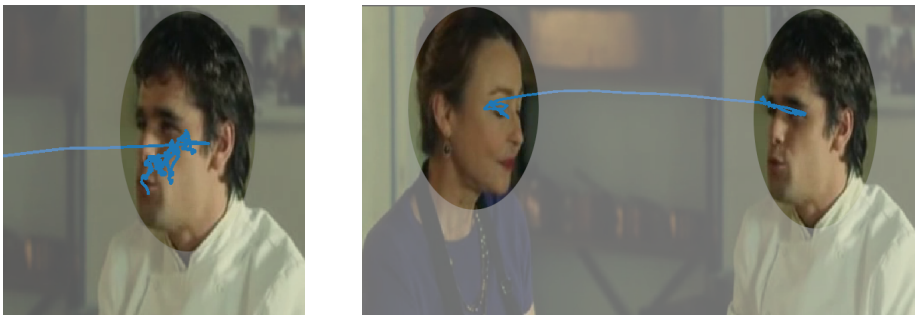
Once the overall support of all patches $\mathcal{M}$ is available, we estimate the parameters defining each patch, namely $\mathcal{P}_p = (\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p, \nu_p)$ representing its location, shape and value respectively. The value is simply inherited from the generating priority map $\nu_p = v_\ell$. Location and shape parameters are derived so to provide an elliptical representation of the patch support (patch centre and axes).

4.5 The attentive stage: stochastic walk driven by audio-visual patches

At this point, the input for the attentive stage is available in the form of value-sensitive foraging patches $\mathcal{P} = \{\mathcal{P}_p(t)\}_{p=1}^{N_P}$, with $\mathcal{P}_p(t) = (\boldsymbol{\mu}_p(t), \boldsymbol{\Sigma}_p(t), \nu_p(t))$, that define the multimodal landscape for the forager's walk.

4.5.1 Dynamics of the walk

Consider the simple case where a single patch of the viewed scene centered at location $\boldsymbol{\mu}$ (center of mass) serves as an attentional attractor, e.g. the face patch in Fig.4.6a. The gaze approximately fluctuates (fixational movement) for a time interval around $\boldsymbol{\mu}$.



(a) **(b)**

Figure 4.6: *(a) A face patch serving as attractor of attention, where the gaze deployment in time can be described as a biased 2-D random walk (b) Two face patches representing multiple centers of attraction, with an example of fixation and relocation among patches*

We can idealise the motion of gaze as that of a particle. In Newtonian dynamics the attraction of a particle of position $\mathbf{r}(t)$ pulled towards the location $\boldsymbol{\mu}$ can be described by means of a potential function, a quadratic form $H(\mathbf{r}, t) = \frac{1}{2}(\boldsymbol{\mu} - \mathbf{r}(t))^T \mathbf{B}(\boldsymbol{\mu} - \mathbf{r}(t))$ that controls the particle's direction and velocity $\dot{\mathbf{r}}(t)$; in particular, $\mathbf{B}$ is the 2×2 matrix that constrains the strength of the attraction. In the case that friction is high, particle's velocity is not directly involved and the equation of motion can be written (Nelson, 1967; Brillinger et al., 2002)

$$d\mathbf{r}_F(t) = -\nabla H(\mathbf{r}_F, t)dt, \tag{4.7}$$

with $\nabla = (\partial/\partial x, \partial/\partial y)^T$ the gradient operator applied to the potential and defining the force field $\mathbf{F} = \nabla H$.

When motion is subject to random forces, Eq. 4.7 generalizes to the stochastic differential equation (SDE)

$$d\mathbf{r}_F(t) = \mathbf{B}[\boldsymbol{\mu} - \mathbf{r}_F(t)]dt + \mathbf{D}(\mathbf{r}_F(t))d\mathbf{W}(t), \tag{4.8}$$

where $\mathbf{B}[\boldsymbol{\mu} - \mathbf{r}_F(t)] = -\nabla H(\mathbf{r}_F, t)$ is the drift term, $\mathbf{D}$ is a 2×2 matrix representing the diffusion parameter. The noise term $\mathbf{W}(t)$ is a 2-D Brownian process that leads to variability around deterministic motion. Simply put, in the stochastic case the particle (gaze) is wandering but being pulled towards the location $\boldsymbol{\mu}$.

Eq. 4.8 can be easily recognised as a Langevin-type equation. Precisely, the gaze trajectory $\mathbf{r}_F(t), t \geq 0$ is an instance of the 2-D mean-reverting Ornstein-Uhlenbeck (OU) process, where typically $\mathbf{B} = (b_x, b_y)^T$, $\mathbf{D}\mathbf{D}^T = \sigma^2\mathbb{I}$ and $\mathbf{W} = (W_x, W_y)^T$ are independent Brownian processes. Clearly, when $\mathbf{B} = \mathbf{0}$, the drift term is $\mathbf{0}$ and the OU process boils down to the Brownian walk.

As pointed out in Section 3.3.3 the general solution for Equation 4.8 writes:

$$\mathbf{r}_F(t') \mid \mathbf{r}_F(t) \sim \mathcal{N}(\boldsymbol{\mu} + e^{-\mathbf{B}\delta t}(\mathbf{r}_F(t) - \boldsymbol{\mu}), \boldsymbol{\Psi}), \tag{4.9}$$

Equation 4.9 describes gaze dynamics towards one point of attraction. In our case, the visual landscape is a time-varying landscape with multiple attractors, the centres of patches $\mathcal{P}_p$. This problem has been partially considered in animal ecology. Breed et al.

(2017) have proposed a multi-state extension of Eq. 4.9, considering multiple centers of attraction.

These centers have unique OU parameters $\boldsymbol{\mu}_i, \mathbf{B}_i, \boldsymbol{\Psi}_i$. However, relocation paths between attractors are not explicitly modelled, which in our case would correspond to the important case of medium/long saccades. Also, along time multimodal patches can vary in number, shape and value.

Harris and Blackwell (2013) proposed a flexible class of continuous-time models for animal movement, allowing movement behaviour to depend on location in terms of a discrete set of regions and also on an underlying behavioural state. The diffusion processes that the individual follows while in a particular combination of state and region are, by assumption, OU processes. Thus, for each combination, the parameters of the OU process are specified as, $\boldsymbol{\mu}_i^{(s)}, \mathbf{B}_i^{(s)}, \boldsymbol{\Psi}_i^{(s)}$, for states $s = 1, ..., K$, and regions $i = 1, ..., L$. The switching process is a continuous- time finite state Markov chain. Its properties are therefore defined by its generator (Harris and Blackwell, 2013), the matrix of instantaneous rates of transition between states observed at short time intervals of length δt. Again, such approach is unfeasible, in our case, where the number of attractors - and, consequently, the number of states- is not known a priori and varies in time.

In our case, we are more truly dealing with two behavioural states that are independent of location: local intensive foraging and extensive exploration. Denote $\{S(t) : t \geqslant 0\}$ a process defined on a binary set $s_t \in \{0, 1\}$ accounting for such behaviour switching process. Its value represents which state of the hidden behaviour is active: foraging, when $s_t = 1$, or exploration when $s_t = 0$ at time t. The regions of attraction are represented by the ensemble of patches $\mathcal{W}(t) = \{\mathcal{P}_p(t)\}_{p=1}^{N_P}$.

In this setting the parameters $\boldsymbol{\mu}_p^{(s_t)}, \mathbf{B}_p^{(s_t)}, \boldsymbol{\Psi}_p^{(s_t)}$ of the OU process are related to a chosen patch p identified through its center location parameter $\boldsymbol{\mu}_p^{(s_t)}$. Meanwhile, the state s_t sampled at time t drives the choice of the appropriate parameters $\mathbf{B}_p^{(s_t)}, \boldsymbol{\Psi}_p^{(s_t)}$.

The specification of parameters constrains the OU process to bias the random walk locally, that is in proximity of the patch located at $\boldsymbol{\mu}_p^{(1)}$; alternatively, $\boldsymbol{\mu}_p^{(0)}$ denotes a patch different from current location, which can be reached through displacements at a larger scale defined by $\mathbf{B}_p^{(0)}, \boldsymbol{\Psi}_p^{(0)}$. This way gaze dynamics is given by the multi-state OU equation

$$d\mathbf{r}_F(t) = \mathbf{B}_p^{(s_t)}[\boldsymbol{\mu}_p^{(s_t)} - \mathbf{r}_F(t)]dt + \mathbf{D}_p^{(s_t)}(\mathbf{r}_F(t))d\mathbf{W}^{(s_t)}(t), \tag{4.10}$$

which is solved by

$$\mathbf{r}(t') \mid \mathbf{r}(t) \sim \quad \mathcal{N}(\boldsymbol{\mu}_p^{(s_t)} + e^{-\mathbf{B}_p^{(s_t)}\delta t}(\mathbf{r}(t) - \boldsymbol{\mu}_p^{(s_t)}), \boldsymbol{\Psi}_p^{(s_t)}).$$

with $\boldsymbol{\Psi}_p^{(s_t)} = \boldsymbol{\Gamma}_p^{(s_t)} - e^{-\mathbf{B}_p^{(s_t)}\delta t}\boldsymbol{\Gamma}_p^{(s_t)}e^{-\mathbf{B}_p^{(s_t)'}\delta t}$. To sum up, gaze dynamics is obtained through the following steps:

1. Sample the behavioural state, based on the current experience of the forager (up to time t, and summarised by parameters $\xi(t)$)

$$s(t) \sim P(\xi(t)) \tag{4.11}$$

2. Sample the patch index

$$p(t) \sim P(\boldsymbol{\pi}(t)) \tag{4.12}$$

with $\boldsymbol{\pi}(t)$ the set of parameters depending on the landscape state, and choose patch $\mathcal{P}_p^{(\ell)}$.

3. Set OU parameters $\boldsymbol{\mu}_p^{(s_t)}, \mathbf{B}_p^{(s_t)}, \boldsymbol{\Psi}_p^{(s_t)}$ and sample the gaze shift $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t')$ via the OU process specified by Eq. 4.11, which is explicitly written as

$$\begin{aligned} x_F(t') \mid x_F(t) &\sim \mathcal{N}(\mu_{x,p}^{(s_t)} + e^{-b_{x,p}^{(s_t)}\delta t}(x_F(t) - \mu_{x,p}^{(s_t)}), \psi_{p,x}^{(s_t)}), \\ y_F(t') \mid y_F(t) &\sim \mathcal{N}(\mu_{y,p}^{(s_t)} + e^{-b_{x,p}^{(s_t)}\delta t}(y_F(t) - \mu_{y,p}^{(s_t)}), \psi_{p,y}^{(s_t)}), \end{aligned} \tag{4.13}$$

with $\psi_{p,x}^{(s_t)} = \gamma_x^{(s_t)}(1 - e^{-2b_{x,p}^{(s_t)}\delta t})$ and $\psi_{p,y}^{(s_t)} = \gamma_y^{(s_t)}(1 - e^{-2b_{y,p}^{(s_t)}\delta t})$.

Regarding the OU parameters the drift terms $b_{x,p}^{(s_t)}$ and $b_{y,p}^{(s_t)}$ are set proportional to the width of the patch p if $s_t = 1$, or proportional to the distance to the arriving patch (d_p), otherwise. The diffusion terms $\gamma_x^{(s_t)}, \gamma_y^{(s_t)}$ is set proportional to the average distance between patches if $s_t = 0$; equal to 1 otherwise.

The steps 1 and 2 behind the choice of the forager's action $\mathcal{A}(t) = \{s(t), p(t)\}$ at time t involve explicit calculation of Eqs. 4.11 and 4.12. These are discussed in the following Section.

4.5.2 Switching behaviour: should I stay or should I go?

Assume that the FOA is located at $\mathbf{r}_F(t)$, within the current patch p, and, for simplicity, that gaze is involved in local patch exploitation. The problem that the perceiver moment to moment has to solve boils down to answering the question: Should I stay or should I go?

In its essence, this is a foraging problem. Indeed, answering such question has long been a fundamental objective in ecology in the endeavour of understanding how animals effectively search for and exploit food patches (MacArthur and Pianka, 1966), and, in particular, how a patch cycle is handled. Consider the environment consisting of a set of discrete patches: a cycle starts when the animal leaves a patch to search for a new one; once a patch has been found, the animal gains energy at a rate that decreases as the food becomes depleted; eventually the animal leaves the patch and a new cycle starts.

A series of optimal foraging theories have been developed in line with this objective (see Stephens (1986), for a review). By assuming that animal activities are optimized to maximize the rate of net energy gain, optimal foraging theories provide testable hypotheses as well as bases for interpreting complex animal behaviour.

Charnov's marginal value theorem (MVT) is central to these theories (Charnov, 1976). The MVT proposes that foragers should exploit patches in such a way as to maximize a net rate of energy gain and predicts the optimal patch residence time. Let G denote the net energy gain on a cycle, and let T denote the time taken to complete a cycle. Simply put, the MVT states that foragers should move from one patch to another when the marginal rate of food intake (thus, of energy gain, $\partial G/\partial T$) drops to the long-term, average rate $\bar{E}$ of food gain across many patches in the environment

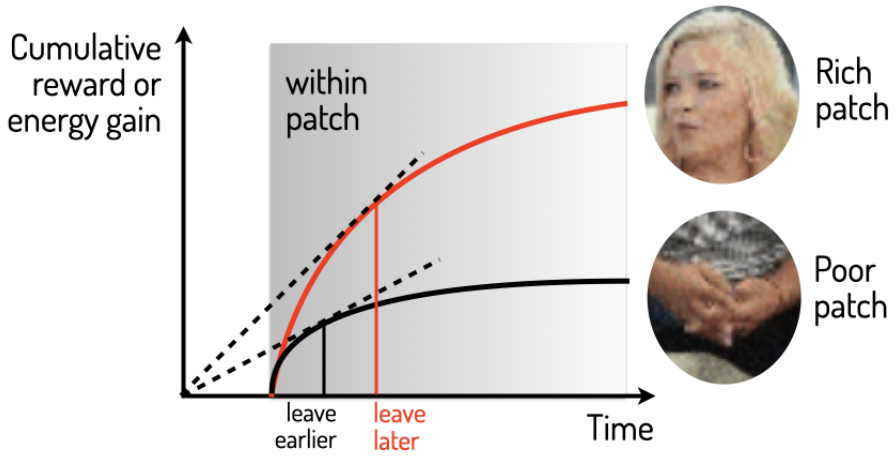


Figure 4.7: *The prediction by MVT is that a poor patch should be abandoned earlier than a rich patch. The time axis starts with a travel time with no energy gain after which the forager finds a patch. The shapes of the red and black gain curves, arising from resource exploitation, represent the cumulative rewards of a "rich" and a "poor" patch, respectively. For each curve, the osculation point of the tangent defines the optimal patch residence time.*

In this simple model, energy gain is a proxy for fitness and it assumes that the foragers have knowledge about the environment: namely, the quality of other patches and traveling time between patches. Thus, MVT predicts that patch quality should affect patch leaving. Accordingly, a poor patch, yielding a lower energy gain, should be abandoned earlier. Clearly, a forager that stays in a patch too long pays an opportunity cost because it wastes time exploiting a depleted patch when fresher patches remain unexploited.

In a stochastic environment, such as that we are dealing with, where rewards are not deterministic and do not arrive in a smooth flow, an optimal forager should reason about the foraging task probabilistically, based on the potential value of the patch with respect to the environment (McNamara, 1982). The optimal leaving time is when the expected rate, not the observed rate, drops below the average for the environment

In stochastic foraging models, typically G and T are random variables whose distribution depends on the behavioral strategy adopted by the foraging animal. In particular, G is a function of the time varying state $U(t)$ experienced by the forager up to time t, $G(U(t))$; for instance, as detailed later, the value $U(t) = u$, might indicate the number k of items/preys "captured" by the forager. The mean net rate of energetic gain, or mean reward rate, achieved by the animal is defined as ratio of expectations $\mathbb{E}[G]/\mathbb{E}[T]$.

In a stochastic perspective, it is convenient to consider the instantaneous reward rate (McNamara, 1982)

$$g(u,t) = \lim_{\delta t \to 0} \frac{\mathbb{E}\left[G(U(t+\delta t)) \mid U(t) = u\right] - G(u)}{\delta t}, \tag{4.14}$$

that is the expected reward over the next interval of time δt; such definition provides the stochastic counterpart of the continuous energy intake rate $\partial G/\partial T$ exploited by the MVT.

The general rule adopted by the forager, while scrutinising a patch, is to leave the patch when

$$g(u,t) \leq Q(t), \tag{4.15}$$

that is when the instantaneous reward rate drops below a "quality" threshold Q, which, in general, depends on the richness of the environment, the distance between patches and possibly other factors (in actual foraging, predation risk, etc.)

There is a number of ways to make concrete the rule given in Eq. 4.15. A method for calculating $g(u,t)$ has been given in Bayesian foraging approaches (Iwasa et al., 1981; Rodríguez-Gironés and Vasquez, 1997).

A straightforward method is the following. Assume that one patch contains a discrete number of items, say m. Let n be the items "consumed" in the time t. Then, the experiential state U is represented by the pair (n,t), $G(U(t)) = G(n,t)$ and $g(u,t) = g(n,t)$. At time t_{W_p} spent within the patch, $k = m - n$ are the items remaining.

When foragers search for food items at random, the time required to find one item is assumed to follow the exponential distribution

$$P(T \in [t, t+\delta t]) = \lambda e^{-\lambda t} dt = Ake^{-Akt} dt, \tag{4.16}$$

where the rate $\lambda = Ak$ depends on A, the searching efficiency of the forager. The probability of capturing at least one item, conditionally on the k remaining, is $P(\delta t \mid k) = 1 - e^{-Akt}$.

It has been calculated (Iwasa et al., 1981; Rodríguez-Gironés and Vasquez, 1997) that, if the initial distribution of the m_p items in patch p (prior, with $k = 0$) follows a Poisson law, $Pois(\rho_p) = \frac{e^{-\rho_p}\rho_p^{m_p}}{m_p!}$, then simply

$$g_p(t_{W_p}) = \rho_p e^{-At_{W_p}}. \tag{4.17}$$

It can be seen from Eq. 4.17 and Eq. 4.16 that the foraging efficiency parameter A controls the rate at which the forager switches from one item to another and consequently the instantaneous intake rate. Yet, it is known that individuals concentrate their foraging effort in areas with high reward (de Knegt et al., 2007; Kazimierski et al., 2016), increasing the handling time of each item, thus increasing the expected time to next item within the patch. In our case, this effect is accounted for by setting $A = \frac{\phi}{\nu_p(t)}$, recalling that $\nu_p(t) \in [0,1]$ is the value associated to the patch p at time t, while ϕ is a positive constant defining the baseline foraging efficiency.

Also, we set ρ as a function of the patch quality, namely,

$$\rho_p(t) = \nu_p(t)|\mathcal{P}_p|e^{-\kappa d_p}, \tag{4.18}$$

where $|\mathcal{P}_p|$, is the area of the patch, ν_p is the patch value, and their product is weighted by $e^{-\kappa d_p}$ representing the visibility of the patch, d_p being the distance to patch p from the current point of gaze and κ being a positive constant. In foraging terms, the weighting factor accounts for the cost of relocating between patches in foraging .

The expected average gain from the environment for all patches q except the current one can be obtained by considering the potential intake rate at $t_W = 0$, i.e., via Eq. 4.17 $g_q(0) = \rho_q$, $q \neq p$:

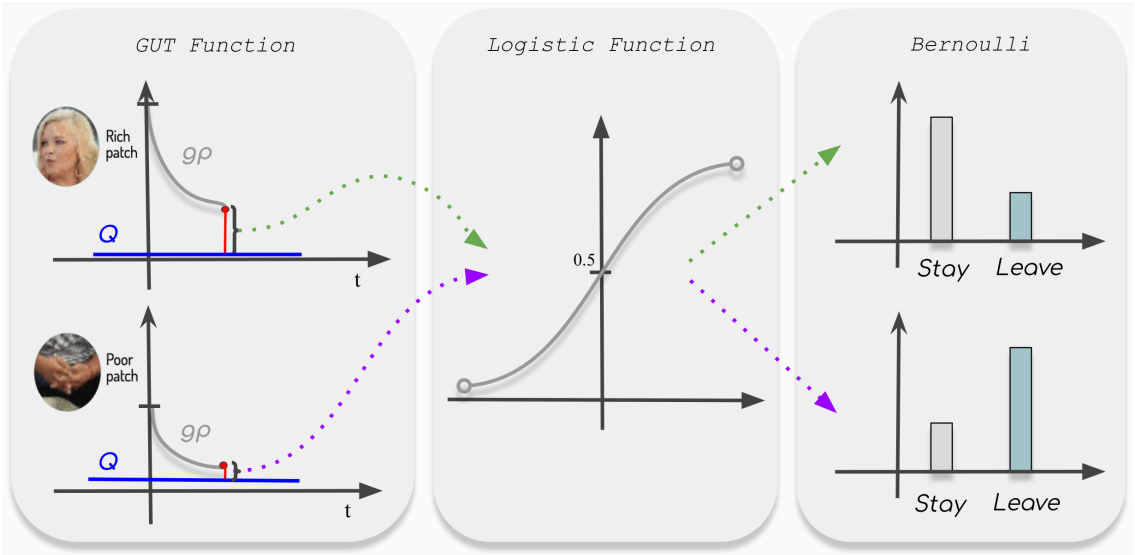


Figure 4.8: *Overall description of the switching behaviour. The first block depicts the typical conduct of the instantaneous reward rate for two types of patches (rich and poor). These can be conceived as Giving Up Time (GUT) functions; as time goes by the GUT function approaches the quality threshold Q, the run being faster for poorer patches. At any time step the decision stay/go is taken by sampling a Bernoulli RV (third block) whose parameter is given by the distance between the GUT function and the quality threshold at that time (opportunely scaled by a logistic function, c.f.r. second block)*

$$Q(t) = \frac{1}{N_p - 1} \sum_{q \neq p} \rho_q(t). \tag{4.19}$$

Rather than straightforwardly use the deterministic rule given in Eq. 4.15, we allow the forager to perform a probabilistic decision; namely the behavioural state decision $s(t) \in 0, 1$ is sampled following a Bernoulli law, $Bern(s(t) \mid \xi(t))$. The parameter ξ, denoting the prior probability of staying within the patch is obtained using a logistic rule accounting for a stochastic comparison on the difference $g_p(t_{W_p}) - Q$, thus

$$\xi(t) = P(stay \mid g(t), Q(t))) = \frac{1}{1 + e^{-\beta(g_p(t_{W_p}) - Q(t))}}, \tag{4.20}$$

$$s(t) \sim Bern(\xi(t)). \tag{4.21}$$

By random sampling the behavioral state $s(t)$, most of the time we are likely to get a state "sample" that is somewhere close to the prior $\xi(t)$. However, sometimes we will randomly sample a decision in the tails of the distribution, which offers an opportunity to the forager to tradeoff between the determinism/trend set by rule given Eq. 4.15, and the dynamically varying landscape. The overall procedure is succinctly depicted in Figure 4.8.

Eventually, if $s(t) = 0$ is sampled, the choice of a patch is the next step to be accomplished.

4.5.3 Choosing the next patch

Given the N_P patches, denote π_p the probability of choosing indexed by patch $p = 1, \cdots, N_P$, with $\sum_p \pi_p = 1$ Then, the sample space of multiple choices can be considered to be the set of 1-of-K encoded (Bishop, 2006) random vectors $\mathbf{c}$ of dimension $K = N_P$ having the property that exactly one element c_p has the value 1 and the others have the value 0. The particular element having the value 1 indicates which patch has been chosen. In other terms, $\mathbf{c}$, follows a categorical (or generalised Bernoulli) distribution, $\mathbf{c} \sim Cat(\boldsymbol{\pi}, N_P) = \prod_{p=1}^{N_p} \pi_p^{y_p}$. Probabilities $\boldsymbol{\pi} = (\pi_1 \cdots \pi_{N_P})$ can be related to the above described patch model as follows.

The N_P patches can be considered at time t as sources of independent Poisson processes $M_p(t) \sim Pois(\rho_p(t))$ with mean value function $\mathbb{E}[M_p(t)] = \rho_p(t)$. Then, in virtue of the superposition theorem , the process $M(t) = \sum_{p=1}^{M_p} M_p(t)$ is a Poisson process with expected value $\mathbb{E}[M(t)] = \sum_{p=1}^{N_p} \rho_p(t) = \rho(t)$.

Under such conditions, the coloring theorem holds , and the vector $(M_1(t)/S, ..., M_{N_p}/S)$, where $S = M_1(t) + ... + M_{N_p}$, follows a multinomial distribution with parameters $\pi_p = \frac{\rho_p(t)}{\rho(t)}$.

When considering a single draw, the multinomial distribution is nothing but the categorical distribution; thus, patch choice can be performed by sampling, at any time t the choice vector

$$\mathbf{c} \sim Cat(\pi_1, \cdots, \pi_{N_P}) = \prod_{p=1}^{N_p} \left[\frac{\rho_p(t)}{\rho(t)} \right]^{c_p}, \tag{4.22}$$

and by selecting patch $\mathcal{P}_p$ based on index p such that $c_p = 1$.

Equation 4.22 together with Eqs. 4.20, 4.21 completely specify Eqs. 4.12 and 4.11, respectively.

4.6 Summary

Attention supports our urge to forage on social cues. Under certain circumstances, we spend the majority of time scrutinising people, markedly their eyes and faces, and spotting persons that are talking. To account for such behaviour, this Chapter develops a computational model for the deployment of gaze within a multimodal landscape, namely a conversational scene. Gaze dynamics is derived in a principled way by reformulating attention deployment as a stochastic foraging problem. An Ornstein-Uhlenbeck process with switching parameters has been employed to model the stochastic dynamics of the eye movements at the microscopic level. The switching signal is provided by a stochastic decision making mechanism derived from Charnov's Marginal Value Theorem.

CHAPTER 5

Simulations and results

THE rationale behind experiments described in this Chapter is to figure out whether behaviours simulated from the proposed model are characterized by statistical properties that are significantly close to those featured by human subjects who have been eye-tracked while watching conversational videos. In simple terms, any model can be considered adequate if model-generated scan paths could have been generated by human observers (which we regard as samples of the `Real` model) while attending to the same audio-visual stimuli.

Consider for example Fig. 5.1. It summarises the essential spatio-temporal features computed from scan paths that have been sampled via the GazeDeploy procedure (Algorithm 1) on one clip; these are compared to those of human observers on the same clip. Notably, such results are by and large representative of those obtained on the whole dataset.

The simulation has generated scan paths that *prima facie* mimick human scan paths in terms of spatio-temporal statistics. The actual saccade amplitude distribution exhibits a multi modal shape, which is well replicated by the saccades distribution obtained from model simulation (Figure 5.1b). The model correctly favors small gaze shifts over large ones, that are occasionally undertaken, as highlighted by the right-skewed, long-tailed shape (Dorr et al., 2005). For what concerns the fixation duration (Fig. 5.1e), again, distributions from both real and simulated data exhibit a right-skewed and heavy-tailed shape. This is important, since in our model duration is closely related to the modelling of patch giving up time. Apparently, a high similarity can be noticed between saccades direction distributions of real (Fig. 5.1c) and simulated data (Fig. 5.1f).

Clearly, beyond the adequate behaviour of the model discernible from such qualitative results, the latter need to be quantitatively substantiated. Are such similarities significant from a statistical standpoint? Is the audio-visual information effectively ex-

ploited? Could a different gaze control algorithm provide comparable or even better results?

There are two critical aspects in answering such question.

The first relates to method comparison. Unfortunately a handful of models have been proposed and are experimentally ready for use (i.e., with released code) for predicting gaze shift dynamics. They are referred to as saccadic models (Le Meur and Liu, 2015) and mostly conceived for processing static image input (Itti et al., 1998; Boccignone and Ferraro, 2004; Le Meur and Liu, 2015; Xia et al., 2019; Xia and Quan, 2020; Bao and Chen, 2020b). Two methods are actually available for handling time-varying stimuli, which we used in our experiment (Boccignone and Ferraro, 2014; Zanca et al., 2019).

The second aspect relates to the evaluation metrics. Unlike to classic work on saliency estimation, where standard metrics are available and widely adopted, here assessment must necessarily involve scan path evaluation. Here we adopt two well known and state of the art methods: the ScanMatch (Cristino et al., 2010) and the MultiMatch (Jarodzka et al., 2010; Dewhurst et al., 2012) metrics. ScanMatch is apt to provide an overall performance summary, whilst MultiMatch specifically addresses the many dimensions of gaze dynamics. The evaluation of metric results is subtle, thus we support it by addressing appropriate statistical analyses, a point that is often neglected in computational modelling of visual attention.

5.1 Stimuli and eye-tracking data

The adopted dataset (Xu et al., 2018) consists of 65 one-shot conversation scenes from YouTube and Youku, involving 1 to 27 different faces for each scene. The duration of the videos is cut down to be around 20 seconds, with a resolution of 1280×720 pixels at a frame rate of 25 fps. The dataset includes eye-tracking recordings from 39 different participants (26 males and 13 females, ageing from 20 to 49), who were not aware of the purpose of the experiment. The eye fixations position and duration of the 39 subjects were recorded by a Tobii X2-60 eye tracker at 60 Hz.

Ten subjects were randomly sampled out of the 39 and their scan paths used to determine the free parameters of the model described in Section 4.5.2, namely the baseline foraging efficiency ϕ, the logistic growth rate β and the steepness of the exponential determining the visibility of patches κ. A grid search maximising metric scores according to the procedure described in the following Section 5.2 yielded as optimal values: $\phi = 3.5$, $\beta = 20$ and $\kappa = 18$.

The remaining 29 subjects were used for evaluation.

5.2 Evaluation protocol

We compare the scan paths simulated from a number of model-based, "artificial" observers to those recorded from human observers. By considering different models, or variants of the same model, we simulate different groups of observers. We address two experiments. The first (Sec. 5.3) evaluates the behaviour of the GazeDeploy procedure (thus, exploiting the gaze control strategy described in Algorithm 1) by inhibiting modules accounting for different levels of preattentive information. This provides a family

of models, that are ablated variants of what we name the `Full` model.

The second experiment (Sec. 5.4) compares the `Full` model with other gaze control strategies.

In both experiments, the evaluation protocol is the following. For each video:

1. Compute MultiMatch and ScanMatch scores for each possible pair of the 29 real observers (`Real` vs. `Real`).
2. For each model:
   (a) Generate gaze trajectories from artificial observers.
   (b) Parse/classify trajectories into scan paths (saccades and fixations with the relative duration) via the NSLR-HMM algorithm Pekkanen and Lappi (2017).
   (c) Compute MultiMatch and ScanMatch scores for each possible pair of real and 29 artificial scan paths (`Real` vs. `Model`).
3. Return the average ScanMatch and MultiMatch scores for `Real` vs. `Real` and `Real` vs. `Model` comparisons.

As to point 2b), note that (cfr. Fig. 4.4) the gaze position sequence sampled by GazeDeploy (and its variants) can be assimilated to gaze *raw data* (continuous gaze trajectories) generated by eye-trackers. Thus, in order to follow a classic eye tracking analysis pipeline, the first step is to apply an *event detection* algorithm to both simulated and actual gaze trajectories so to derive the corresponding scan paths (a sequence of fixations). We rely on the NSLR-HMM algorithm described in Pekkanen and Lappi (2017) which classifies the raw data into saccades, fixations, smooth pursuits, and post-saccadic oscillations (PSO). Here we are dealing with dynamic stimuli that can trigger smooth pursuit eye movements. Yet, pursuit can be broadly categorised as a prolonged fixation on a moving target; consequently, smooth pursuit and fixations are collapsed into a single class. Also, saccades embed PSOs.

In what follows we treat each MultiMatch dimension as a stand-alone score. Thus, the analysis uses six different scores: the five obtained from the MultiMatch (MM) dimensions of shape (MM_{Shape}), direction (MM_{Dir}), length (MM_{Len}), position (MM_{Pos}) and duration (MM_{Dur}), plus the ScanMatch score SM.

5.3 Information level effects: the model under the knife

A basic assumption of the proposed model (A3, Section 4.3), derived from psychological studies (Foulsham et al., 2010), states that in a scene displaying conversations and social interactions, attention is predominantly allocated to faces, with higher relevance given to speakers. This assumption is practically addressed in the model by relying on audio-visual priority maps of speaker vs. non-speakers maps.

If such premise holds, we expect that the "ablation" of model components accounting for face information and specifically for speaker information would lead the model-generated scan paths to significantly deviate, in a statistical sense, from human scan paths.

On the other hand, given that the availability of such information is necessary for a human-like gaze deployment, is it sufficient? To put the question straight: when

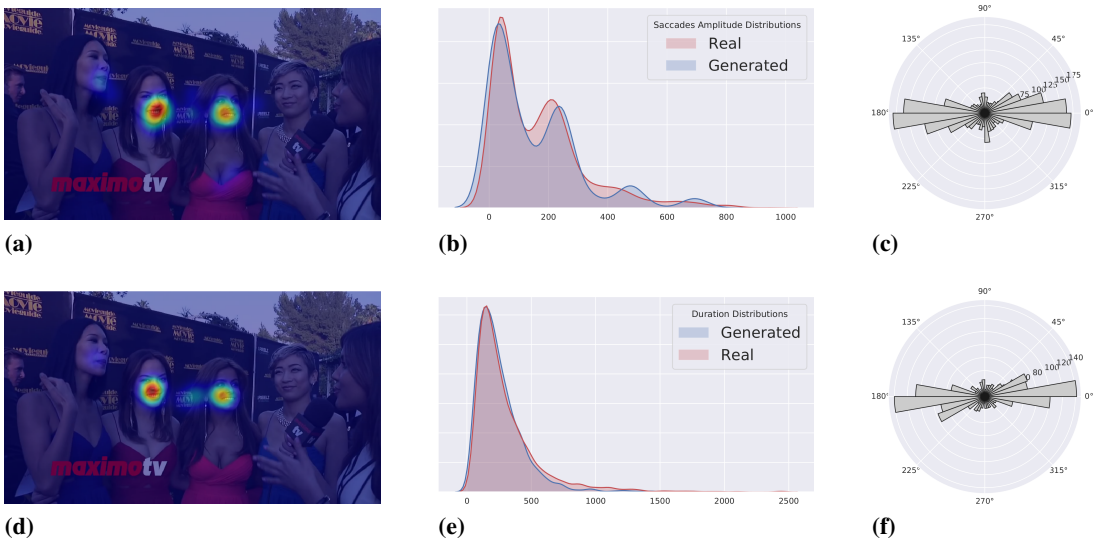


Figure 5.1: *(a) Frame of video* 010 *with overlaid heatmap of real fixations. (b) Real (red) and Generated (blue) saccades amplitude distribution. (c) Real saccades direction distribution. (d) Frame of video* 010 *with overlaid heatmap of generated fixations. (e) Real (red) and Generated (blue) fixations duration distribution. (f) Generated saccades direction distribution.*

attending a conversational clip, do we actually need bottom-up information/saliency for reliably generating gaze shifts, or is it redundant? This is a deceptively simple point that has been overlooked, since by and large visual attention models give for granted a central role for low-level salience.

In order to shed light on such questions we simulate gaze data from the following models:

1. `BU` or `Bottom-up`: we prevent the model from the computation of the audio-visual priority maps, thus only considering low-level features in the preattentive stage;

2. `BU+F` or `No Speaker`: faces are considered, together with `BU` features, but no distinction is made between speakers and non-speakers;

3. `F` or `Face` model: only faces are considered, as in the `No Speaker` model, but without `BU` features;

4. `F+S` model: a face and speaker model, without `BU` features;

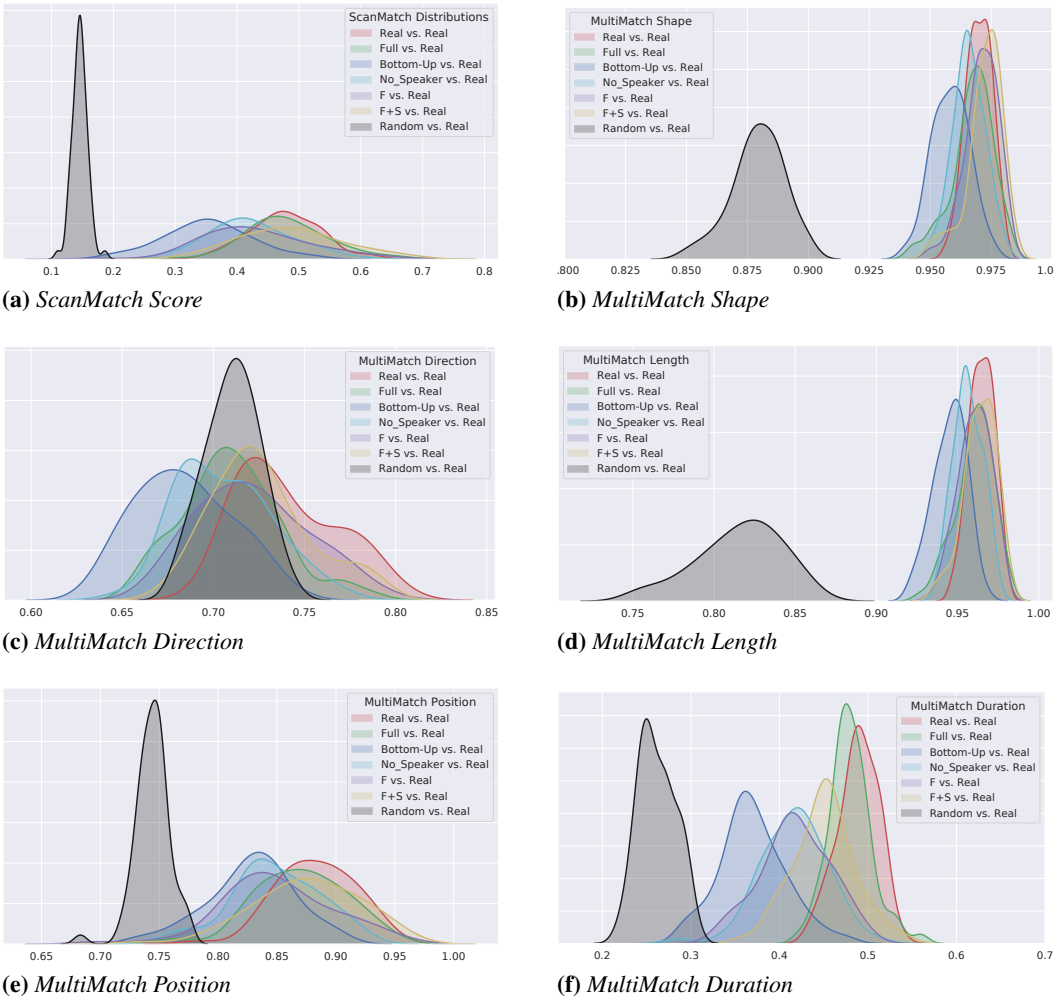
5. `BU+F+S` or `Full`: the model described in this paper where audio-visual patches account for low-level information (`BU`), faces (`F`) and speakers (`S`).

In addition, a baseline `Random` model is adopted, too. This simply generates random gaze shifts by sampling (x, y) fixation coordinates and relative duration from the uniform distribution. Note that in such setting, only the `Full` and the `F+S` models are explicitly accounting for audio information.

For each model we adopt the protocol described in Section 5.2. Figure 5.2, depicts at a glance the empirical distributions of the scores obtained in the ablation experiments. A preliminary inspection shows that the `Full` and `F+S` models give rise to distributions that are close to those yielded by real subjects for all dimensions, with the exception of the direction score MM_{Dir}.

5.3.1 Statistical analyses

The similarity scores obtained from the six models introduced above are used to assess whether or not a model generates scan paths that significantly differ from those of human observers and to gauge the size of such difference (effect size). In the analyses that follow, scores obtained from `Real` vs. `Real` comparison represent the gold standard; the significance level of all statistical tests is $\alpha = 0.05$.



(a) *ScanMatch Score*

(b) *MultiMatch Shape*

(c) *MultiMatch Direction*

(d) *MultiMatch Length*

(e) *MultiMatch Position*

(f) *MultiMatch Duration*

Figure 5.2: *Score distributions for models considered in the ablation experiment*

For each score, the normality of model distributions (groups) was assessed via the Shapiro-Wilk test for normality (Shapiro and Wilk, 1965) with Bonferroni correction (Bonferroni, 1936). All models exhibit normal distributions for scores SM, MM_{Len}

	M	SD	d	Magnitude
F+S	0.487	0.082	-0.127	negligible
Full	0.481	0.067	-0.045	negligible
Face	0.430	0.088	0.650	medium
No Speaker	0.423	0.067	0.889	large
Bottom-Up	0.352	0.071	1.955	large
Random	0.146	0.012	8.140	large
Real	0.478	0.056	0	/

(a) *ScanMatch Score*

	MED	MAD	δ	Magnitude
F+S	0.974	0.006	-0.300	small
Face	0.971	0.007	-0.140	negligible
Full	0.969	0.007	0.170	small
No Speaker	0.965	0.006	0.391	medium
Bottom-Up	0.959	0.009	0.790	large
Random	0.880	0.010	1.000	large
Real	0.970	0.005	0	/

(b) *MultiMatch Shape*

	MED	MAD	δ	Magnitude
F+S	0.722	0.022	0.283	small
Face	0.718	0.030	0.351	medium
Random	0.711	0.016	0.651	large
Full	0.707	0.024	0.574	large
No Speaker	0.708	0.030	0.628	large
Bottom-Up	0.680	0.024	0.862	large
Real	0.734	0.029	0	/

(c) *MultiMatch Direction*

	M	SD	d	Magnitude
F+S	0.964	0.010	0.116	negligible
Full	0.960	0.011	0.462	small
Face	0.961	0.010	0.486	small
No_Speaker	0.957	0.009	1.034	large
Bottom-Up	0.945	0.010	2.186	large
Random	0.816	0.026	7.726	large
Real	0.965	0.007	0	/

(d) *MultiMatch Length*

	MED	MAD	δ	Magnitude
F+S	0.878	0.048	0.093	negligible
Full	0.869	0.044	0.205	small
Face	0.841	0.040	0.459	medium
No Speaker	0.844	0.035	0.516	large
Bottom-Up	0.830	0.034	0.748	large
Random	0.745	0.012	1.000	large
Real	0.885	0.038	0	/

(e) *MultiMatch Position*

	M	SD	d	Magnitude
Full	0.480	0.024	0.384	small
F+S	0.450	0.035	1.328	large
Face	0.418	0.038	2.269	large
No Speaker	0.416	0.037	2.355	large
Bottom-Up	0.370	0.037	3.866	large
Random	0.262	0.021	10.603	large
Real	0.489	0.022	0	/

(f) *MultiMatch Duration*

Table 5.1: ***Information level effects****: central tendencies for each score and model computed as mean (**M**) or median (**MED**) with associated dispersion metrics (standard deviation, **SD** or median absolute deviation, **MAD**. Effect sizes are computed as the Cohen's* $\mathbf{d}$ *or the Cliff's* $\boldsymbol{\delta}$ *between the given model and real subjects.*

and MM_{Dur}; when MM_{Shape}, MM_{Dir} and MM_{Pos} scores were considered, the null hypothesis of normality was rejected for at least one of the models.

Then, for normally distributed scores the statistics adopted to summarize each model were the empirical mean and standard deviation. The effect size for each model was measured via Cohen's d (Cohen, 2013), based on differences between model and `Real` means. Otherwise, if at least one model violated normality, we considered the median for capturing the central tendency and the absolute deviation from the median as the dispersion measure. In that case the effect size for each model was computed via Cliff's delta (Cliff, 2014).

The overall results are reported in Table 5.1. We follow Cohen's convention (Cohen, 2013) considering effect magnitudes 'small' ($d \sim 0.2$), 'medium' ($d \sim 0.5$), 'large' ($d \sim 0.8$) and negligible ($d < 0.2$). As to Cliff's delta, we follow Hess and Kromrey (2004), by distinguishing 'small' ($\delta \sim 0.147$), 'medium' ($\delta \sim 0.33$) and 'large' ($\delta \sim 0.474$) effect size; the effect is negligible for $\delta < 0.147$.

We then performed homogeneity of variance tests. For each score, when normality held, the Bartlett's test (Bartlett, 1937) was employed to test homoscedasticity; otherwise, Levene's test was adopted (Levene, 1961). Either Bartlett's or Levene's tests rejected the null hypothesis of homogeneity of variances ($p < 0.01$, for all scores).

The assessment of statistically significant differences between models was performed as follows. Since neither normality, nor equality of variances could be ensured,

we resorted to the well known Friedman Test (FT, Friedman (1937), a non-parametric variant of ANOVA), with Nemenyi Nemenyi (1963) *post-hoc* analysis of pairwise differences (similar to the Tukey test for ANOVA). We tested the null hypothesis for each score that the medians were equal between the 6 groups plus the `Random` one.

For all scores, the FT rejected the null hypothesis ($p < 0.001$, always, cfr. Fig. 5.3). Thus, for each score at least one statistically significant difference between two models exists. The Nemenyi's *post-hoc* analysis was then performed. The test compares each pair of groups in terms of their difference in average ranks; if such difference exceeds the critical difference CD_α at the confidence level α, then the two group are statistically different. Figure 5.3 reports the FT outcomes (test statistics t and p-value p) and, most important, visualises *post-hoc* analysis results. The latter are rendered in a compact, information-dense format by means of the *Critical Difference (CD) Diagram* as proposed in Demšar (2006). CD Diagrams show the average rank of each model (higher ranks meaning higher average scores); models whose difference in ranks does not exceed the CD_α ($\alpha = 0.05$) are joined by thick lines and cannot be considered significantly different.

By first considering the ScanMatch metric, a clear ranking is established. We can assume that there are no significant differences within the following two groups: `F+S`, `Real` and `Full`; `Face` and `No_Speaker`. All other differences are significant. Taking into account the magnitude of the effect, the difference between the `Full` model and `Real` is negligible with a smaller magnitude than that of `F+S`. The effect size grows to large for `Face` and `No_Speaker`. The `Bottom-Up` model performs badly, albeit being significantly different from the `Random` model, which clearly has the largest effect size.

Together with the fact that when the `BU` component is ablated from higher level models, the similarity performance does not decrease, these results suggest that `BU` conspicuity has a modest relevance, at least for the kind of conversational clips we deal with.

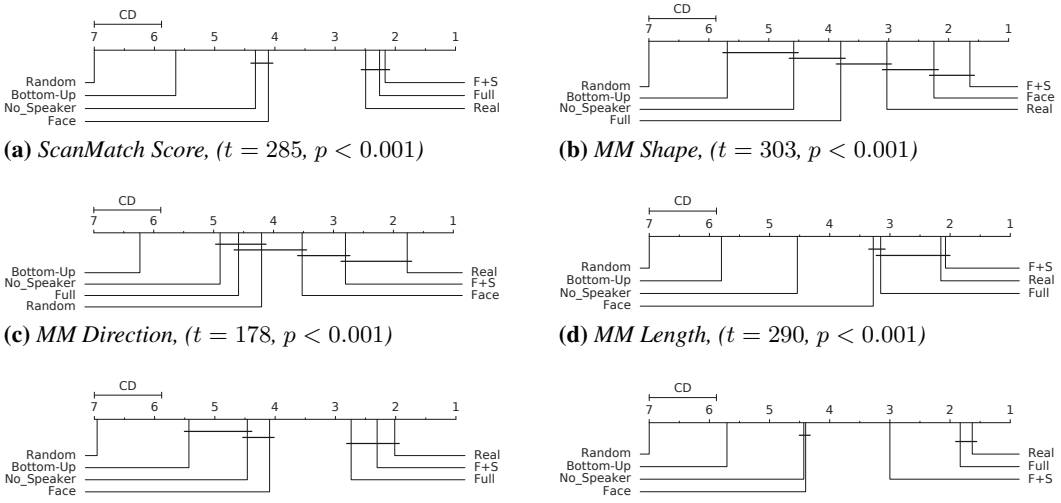
Overall, it can be noted that the test does not support any statistically significant difference between the scores of `Real` subjects and the ones from the `Full` model. This is true for the ScanMatch metric and for all the MultiMatch metric dimensions except for the *Direction* score. A similar behaviour is exhibited by the `F+S` model, the only remarkable difference being that of the MultiMatch *Duration* metric. In this case, as opposed to the `Full` model a significant difference with fixation duration of real subjects is found. Significant differences arise when comparing real scan paths with those generated from ablated models like the `Bottom-Up`, `No Speaker` and `Face` and the huge dissimilarity with the randomly generated eye movements (`Random` model).

Taken together with the size effects reported in Table 5.1, these results bear some consequences. First of all, they show how the proposed (`Full`) model is able to mimic the human behaviour of gaze deployment to audio-visual dynamic stimuli of social interactions. This is witnessed by differences with the scores achieved by real subjects that are negligible in their size and not statistically significant for almost all scores. The only exception is the MultiMatch *Direction* dimension, for which no clear association with the *gold standard* is found. This is not surprising. Indeed the saccades direction is the only feature that is not explicitly tackled in any aspect of the proposed model, but only subsumed as a consequence of the value based patch selection mechanism (Eq.

4.22).

Second, it is interesting to note how preventing models from accounting for bottom-up information, does not results in a significant loss of performance, according to most of the adopted metrics, when comparing with the same models that account for it. Indeed, if fixation duration seems to benefit from the computation of low level cues, for other scores like the *ScanMatch* and the MultiMatch *Position*, the ablation of bottom-up information generated outcomes that are indistinguishable from a statistical standpoint. In light of this result, it is clear how the role of bottom-up information when dealing with videos of social interaction, should be reappraised, since marginally contributing to the process of attention allocation.

Overall, the only model that performs comparably with humans is the `Full` model; indeed it is able to achieve indistinguishable results w.r.t. humans on 5 out of the 6 adopted metrics. The fact that the models obtained after the ablation of high level information (speaker/no-speaker, face location) produce significantly lower scores, highlights the causal effect of the presence of (talking) faces, or more generally top down cues, on attention allocation. This fact has been previously demonstrated in psychological studies (Foulsham et al., 2010), but here it is made operational by means of a computational model.



(a) *ScanMatch Score, ($t = 285$, $p < 0.001$)*

(b) *MM Shape, ($t = 303$, $p < 0.001$)*

(c) *MM Direction, ($t = 178$, $p < 0.001$)*

(d) *MM Length, ($t = 290$, $p < 0.001$)*

(e) *MM Position, ($t = 270$, $p < 0.001$)*

(f) *MM Duration, ($t = 328$, $p < 0.001$)*

Figure 5.3: ***Information level effects****: critical Difference (CD) diagrams of the post-hoc Nemenyi test ($\alpha = 0.05$) for the ScanMatch score and each MultiMatch score obtained by using different information levels obtained by ablation of components feeding the GazeDeploy strategy. Diagrams can be read as follows: the difference between two models is significant if the gap between their ranks is larger than CD; there is a line between two models if the rank gap between them is smaller than CD. Graphically, models that are not significantly different from one another are connected by a black CD line. Friedman's test statistic (t) and p-value (p) are reported in brackets.*

5.4 Gaze control effects

The experiment reported in this section aimed at comparing the GazeDeploy control strategy to those of models previously proposed in the literature that are:

- capable of handling time-varying scenes
- for which a model implementation is available

In particular we used the Ecological sampling model (from now on `Eco_Sampling`)[1] proposed in Boccignone and Ferraro (2014), and the recent `G-Eymol` model[2] (Zanca et al., 2019) described in Section 2.3.4.

For what concerns the `Eco_Sampling` model it's worth noticing how, much like GazeDeploy, it assumes the gaze sequence to be generated by a stochastic process. Different from GazeDeploy it does not rely on a specific account for patch handling and giving-up time. Moreover, the preattentive representation, formalised in terms of proto-objects, roughly corresponds to the patches of the GazeDeploy procedure. In what follows, we feed the `Eco_Sampling` model with the same perception of the world as inferred in the preattentive stage of the proposed `Full` model, so as to focus on the performance of the different gaze control strategies, rather than representation issues.

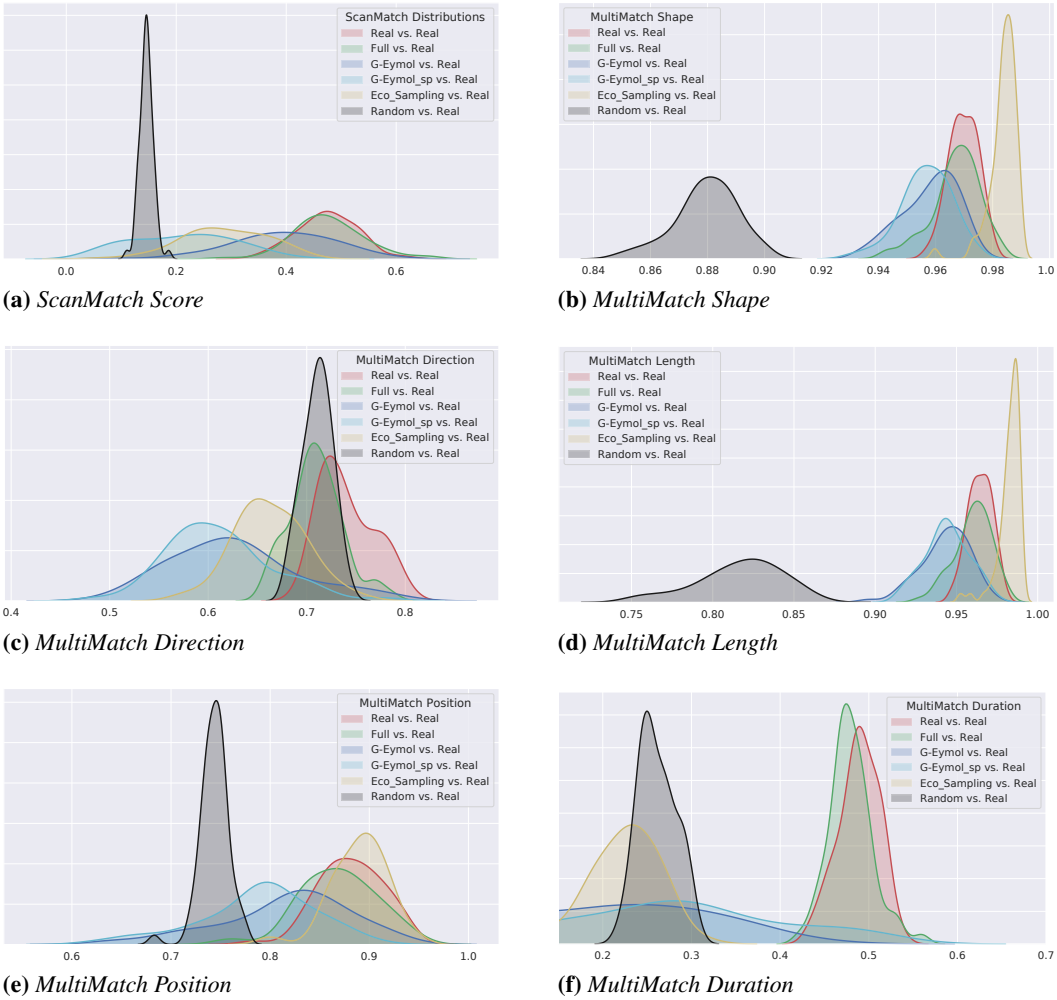
For what concerns the `G-Eymol` model, we employ the version that allows faces as additional masses; this is accomplished in the original implementation, by adopting a Haar cascade face detection (Viola and Jones, 2004).

Further, in order to belay a fair comparison, we set up a variant (`G-Eymol_sp`) that takes into account the difference between speakers and non-speakers. This is achieved by feeding the `G-Eymol` model with speaker and non-speaker masses whose magnitude is proportional to their value as defined in Equation 4.6. The `G-Eymol` equation of motion are deterministic. However, the stochasticity requested to sample different scan paths mimicking different observers can be achieved by perturbing the initial conditions of the equations. Eventually, we also consider the `Random` model.

As in the previous experiment, for each model we adopted the protocol described in Section 5.2. Figure 5.4, depicts at a glance the empirical distributions of the scores obtained by the 5 control models. Visual inspection of distributions derived from the ScanMatch score suggests a higher similarity of scan paths simulated from the `Full` model with respect to the original `G-Eymol`; `Eco_Sampling` and, surprisingly, `G-Eymol_sp` achieve inferior performance.

As to MultiMatch, the behaviour of the `Full` model gives rise to distributions that on the average are closer than other models to those yielded by real subjects, the MM direction score again being an exception, as in the previous experiment. But here, remarkably, the `Full` model seems to outperform the others with respect to the Duration dimension, a result that was to be expected, because this dimension benefits from the Bayesian stochastic foraging approach.

[1] Matlab implementation available at `https://github.com/phuselab/EcoSampling`
[2] Python implementation available at `https://github.com/dariozanca/G-Eymol`



(a) *ScanMatch Score*

(b) *MultiMatch Shape*

(c) *MultiMatch Direction*

(d) *MultiMatch Length*

(e) *MultiMatch Position*

(f) *MultiMatch Duration*

Figure 5.4: *Score distributions for models considered in the gaze control experiment*

5.4.1 Statistical analyses

The statistical analysis of effects entailed by different gaze control strategies, closely followed the one carried out in Section 5.3.1.

Scores SM, MM_{Dir}, MM_{Pos} and MM_{Dur} according to the Shapiro-Wilk test failed to reject the hypothesis of normality, as opposed to scores MM_{Shape} and MM_{Len}. The overall results for central tendencies, dispersions and effect size are reported in Table 5.2. For all scores, either Bartlett's or Levene's tests rejected the hypothesis of homoscedasticity of distributions. Thus, the FT with Nemenyi *post-hoc* analysis was performed. The final results are reported in Figure 5.5.The quantitative results overall support what surmised so far by visually inspecting the score distributions.

Based on the *post-hoc* Nemenyi test, and considering the ScanMatch metric, we assume that there are no significant differences within the following three groups: `Full`, `Real`; `Eco_Sampling` and `G-Eymol_sp`; `G-Eymol_sp` and `Random`. All other differences are significant. The effect size of such differences with respect to the gold standard of human observers can be appreciated in Table 5.2.

	M	SD	d	Magnitude
Full	0.477	0.065	-0.044	negligible
G-Eymol	0.393	0.098	1.031	large
Eco_Sampling	0.288	0.078	2.779	large
G-Eymol_sp	0.212	0.093	3.441	large
Random	0.146	0.012	8.393	large
Real	0.475	0.054	0	/

(a) *ScanMatch Score*

	MED	MAD	δ	Magnitude
Eco_Sampling	0.985	0.003	-0.939	large
Full	0.968	0.007	0.171	small
G-Eymol_sp	0.957	0.008	0.820	large
G-Eymol	0.960	0.008	0.714	large
Random	0.881	0.010	1.000	large
Real	0.969	0.006	0	/

(b) *MultiMatch Shape*

	M	SD	d	Magnitude
Random	0.711	0.015	1.345	large
Full	0.710	0.027	1.131	large
Eco_Sampling	0.663	0.036	2.424	large
G-Eymol	0.626	0.064	2.343	large
G-Eymol_sp	0.610	0.053	3.065	large
Real	0.741	0.027	0	/

(c) *MultiMatch Direction*

	MED	MAD	δ	Magnitude
Eco_Sampling	0.985	0.004	-0.913	large
Full	0.960	0.011	0.220	small
G-Eymol_sp	0.944	0.013	0.803	large
G-Eymol	0.944	0.012	0.810	large
Random	0.820	0.024	1.000	large
Real	0.964	0.009	0	/

(d) *MultiMatch Length*

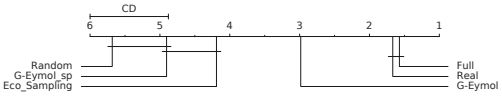
	M	SD	d	Magnitude
Eco_Sampling	0.890	0.028	-0.285	small
Full	0.868	0.039	0.364	small
G-Eymol	0.816	0.066	1.273	large
G-Eymol_sp	0.787	0.059	1.991	large
Random	0.744	0.014	5.549	large
Real	0.881	0.032	0	/

(e) *MultiMatch Position*

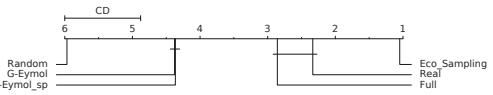
	M	SD	d	Magnitude
Full	0.480	0.024	0.389	small
G-Eymol_sp	0.278	0.122	2.411	large
Random	0.261	0.021	10.562	large
G-Eymol	0.213	0.107	3.556	large
Eco_Sampling	0.221	0.044	7.677	large
Real	0.489	0.022	0	/

(f) *MultiMatch Duration*

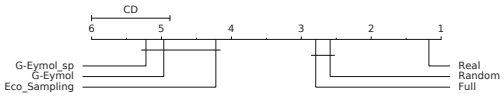
Table 5.2: ***Gaze control effects*** *(notation follows Table 5.1)*



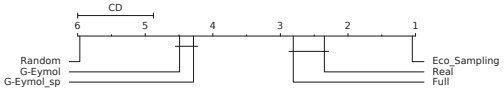
(a) *ScanMatch Score* ($t = 262$, $p < 0.001$)



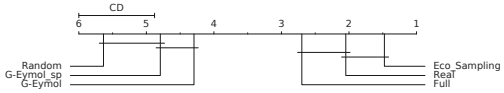
(b) *MM Shape,* ($t = 273$, $p < 0.001$)



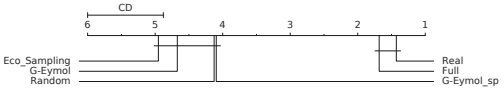
(c) *MM Direction,* ($t = 214$, $p < 0.001$)



(d) *MM Length,* ($t = 274$, $p < 0.001$)



(e) *MM Position,* ($t = 243$, $p < 0.001$)



(f) *MM Duration,* ($t = 206$, $p < 0.001$)

Figure 5.5: ***Gaze control effects****: CD Diagrams of the post-hoc Nemenyi test* ($\alpha = 0.05$) *for MultiMatch (MM) and ScanMatch scores (cfr. Fig 5.3), obtained by using different gaze control strategies (see text for explanation). Friedman's test statistic* (t) *and p-value* (p) *are reported in brackets.*

The many facets of such overall model performance ranking can be best weighed by considering the individual dimensions provided by MultiMatch scores. Remarkable is the divergence with respect to duration: no significant differences are detected within the `Full`and `Real` group, with a small effect magnitude; all other models fall in one group, together with the `Random` model, albeit distinguished by different effect sizes (in this case, for instance `G-Eymol_sp` performs better than the original `G-Eymol`).

The worst performance of all models is detectable on the direction dimension. In this case there are no significant differences within the following two groups: `Full` and `Random` (but with smaller effect size for the first model); `Eco_Sampling`, `G-Eymol` and `G-Eymol_sp`. By taking into account the effect size, models in the second group perform worse than random choice. In simple terms, a random choice of direction seems to provide a better opportunity than any inappropriate strategy.

As to the position dimension there are no significant differences within the following groups: `Eco_Sampling` and `Real`; `Full` and `Real`; `G-Eymol` and `G-Eymol_sp`; `G-Eymol_sp` and `Random`. The smallest effect size in difference from human subjects are provided by the `Full` and `Eco_Sampling` models, the latter being the smallest. This could be due to the fact that, all being equal, the sampling mechanism of interest points from proto-objects/patches can provide a fine-grained, shape-sensitive choice of the possible gaze shift "landing", as opposed to the use of the center of mass attraction in GazeDeploy. Similarly, the high ranking of `Eco_Sampling` is likely to stem from the core business of the approach, namely the sophisticated modelling of gaze shift amplitudes via Lévy flights; yet, GazeDeploy suffers from a smaller effect size. Results achieved for the shape score deserve similar considerations.

5.5 Summary

This Chapter presents simulation experiments from the proposed computational model on a publicly available dataset of eye-tracked subjects. Results show that the simulated scan paths exhibit similar trends of eye movements of human observers watching and listening to conversational clips in a free-viewing condition.

Two experiments were carried over: in the first one we question the usefulness of the adopted modelling assumptions by performing an ablation study of the model. In the second we compare the proposed gaze control strategy with those of state of the art models suggested in literature.

Rigorous statistical analyses proved the soundness of the hypotheses advanced in the previous Chapter and the tight similarity between the generated and real scanpaths, even when comparing with state of the art models.

CHAPTER 6

Discussion and Conclusions

THE present work dealt with the problem of gaze deployment to multi-modal and time varying stimuli. The formalization of such a complex system is not an easy task, especially when considered in it's entirety. In **Chapter 2** it has been shown how often, the most recent attempts to model visual attention do not take into account many of the peculiarities of our oculomotor system; from the dynamics of the attentive process, to the oculomotor biases, up to the inherent stochasticity of gaze allocation.

The main contributions of Chapter 2 is given by the definition of a novel model of time-aware scanpath simulation on static stimuli, which shows how taking into account visual attention dynamics leads to generated scanpaths exhibiting a tighter similarity to the recorded ones, if compared to those generated by time-agnostic models.

Moreover, we stressed on the fact that visual attention is composed by both overt and covert mechanisms; when considered together, both the goals of the observer and the actual gaze shift mechanism must be taken into account. Under such perspective, the observer (either consciously or unconsciously) uninterruptedly has to decide to keep looking at the current portion of the stimuli or relocate gaze elsewhere. Such choices are strongly influenced by the value (quality) that the observer assigns to the current patch. This way of behaving can be successfully associated to that of foraging animals.

The idea of applying movement ecology models to describe visual attention allocation is not new. In **Chapter 3** we have shown how Lèvy Flights may be conceived as an efficient strategy to scan the scene motivated by evolution. Interestingly enough, they represent the missing link between eye movement and foraging models and provide an interesting new perspective on modelling overt attention.

Likewise, Marginal Value Theorem has been employed in the visual attention realm to describe visual search tasks and in particular to address patch leaving mechanisms. This entails a decision making process that deals with the covert attentive mechanisms

that allow to assign specific values to each different patch, based on the current tasks and goals.

These concepts are expanded and made operational in **Chapter 4** where a full computational model of gaze deployment to audio-visual cues of social interactions is proposed.

The main contributions of Chapter 4 lie in the following:

1. The proposed attention deployment model addresses the active sensing of a multimodal stimulus (audio and visual). Although humans are multi-sensory perceivers, surprisingly enough and to the best of our knowledge, there is not much tradition in the computational modelling of this problem.

2. Attention deployment is reformulated as a stochastic foraging problem. Albeit unconventional, this choice allows a parsimonious approach to cope with both the *what* and *how* problems that ground active sensing, the *how* problem being hitherto neglected in the computational modelling of attention.

3. Gaze dynamics succintly relies upon one and only OU stochastic process that is apt to switch between different scales of diffusion. This solution accounts for the variability problem of the perceivers in a simpler way than some attempts based on more cumbersome mathematical tools (e.g., Lévy flights). A side consequence is to allow a concrete step towards the unified modelling of different kinds of gaze shifts, a recent trend in eye movement research.

4. The foraging framework is exploited for a seamless but principled integration of attentional control mechanisms that are modulated by value and rewards. In particular it is shown how implicit social reward as elicited by multimodal conversational clips can be inferred and exploited in the loop.

5. Different from the current propensity towards end-to-end approaches, the model-based behavior of gaze deployment provides an explainable account. This is an important feature if the approach is to be used in a subject's mining context (for example, inferring socially-aware psychological traits of the perceiver or atypical development in the appraisal of social cues)

In a nutshell the model aims at answering fundamental questions on the attentive behaviour of a subject who scrutinises and forages on other subjects involved in social interactions, such as:

- *What* defines a patch of audio-visual information valuable to spot?
- *How* is gaze guided within and between patches?

Surprisingly the study of this problem is still in its infancy in the field of computational modelling of visual attention (Rubo and Gamer, 2018; Nguyen et al., 2018; Bylinskii et al., 2016). This state of affairs is in striking contrast with the exponentially spreading body of audio-visual data that convey social content and the need of analyzing the perceiver's behaviour under such circumstances. Unwisely, a large amount of research effort of the computer vision community in the last two decades has by and

large focused on salience estimation from natural scenes, mostly neglecting the dynamics of actual attention deployment, as instantiated by gaze shifts. The shortcomings of this effort become dauntingly palpable when dealing with scenes endowed with rich semantics, where gaze sampling is affected by goals, rewards, social traits and even expectations about future events.

In this respect, it's not surprising that the *How* problem in attention modelling has been more often addressed in computational neuroscience, psychology and robotics fields, where explainability on the one hand and embodiment/simulation purposes on the other explicitly require answering such question. In particular, active inference (Friston et al., 2006) has been adopted as a scheme for describing visual searches and scene construction (Mirza et al., 2016) or action perception (Donnarumma et al., 2017). Similarly, attention allocation has been framed as an active perception problem in which an agent (robot) actively samples the surrounding environment in order to solve a given task, like action/event recognition or target localization (Ognibene et al., 2013; Ognibene and Baldassare, 2014; Ognibene and Demiris, 2013; Lee et al., 2015). In this vein, reinforcement learning has been adopted to model eye-movements during search (Butko and Movellan, 2008) or object recognition tasks (Paletta et al., 2005). Interestingly enough, it has been shown that equipping an agent with an active vision system can enhance the performance and efficiency on visual tasks such as classification (de Croon et al., 2009).

Still and all, these approaches typically put the accent specifically on the problem of saccadic target selection under a given task or goal (Parr and Friston, 2018), while leaving on the background other important facets of gaze deployment, like the description of the mechanics of oculomotion or the temporal modelling of the decision making (fixation duration). This latter aspect is deemed particularly relevant in the psychological field, as the fixation duration speaks for the moment-to-moment information priorities of the visual system, thus potentially unraveling the strategies employed by the brain to serve ongoing behavior (Tatler et al., 2017). As a matter of fact, the vast majority of computational models that address the *How* question remain silent on the determinants of fixation duration (Borji and Itti, 2012), with few notable exceptions (Tatler et al., 2017; Unema et al., 2007; Findlay and Walker, 1999; Nuthmann et al., 2010; Mackay et al., 2012).

Here we deliberately made a fresh step forward in such direction. It has been shown how oculomotor behaviour while attending at social interactions can be effectively and holistically described in terms of the principles of Optimal Foraging Theory.

Gaze dynamics has been derived in a principled way by reformulating attention deployment as a stochastic foraging problem: the perceiver allocates gaze to audio-visual patches much like a forager visits patches in the environment to obtain nourishment. Our model is that of a stochastic forager performing an Ornstein-Uhlenbeck walk by switching to the appropriate scale for engaging in either within-patch exploitation and large between-patch relocations. The foraging dynamics is thus driven by the audio-visual patches that at any time appear relevant as to social value (rewarding). Patches are sampled from spatially-based probabilistic priority maps. These, in turn, are derived by adapting to our framework recent results gained by deep network techniques (Chung and Zisserman, 2016, 2017), so to account for the visual and auditory objects across different analysis scales. Moment to moment, patch value dynamics is inferred

on a video clip, with dynamics parameters being derived on the basis of eye-tracked gaze allocation of a number of actual observers. Patch choice, handling and leave are framed within an optimal Bayesian foraging setting.

In **Chapter 5** model simulation experiments on a publicly available dataset of eye-tracked subjects and in-depth statistical analyses of results so far achieved have been performed. These show an overall statistically significant similarity between scan paths of human observers and those generated by the GazeDeploy procedure, which uses the full stack of information levels.

The current model has limitations and caveats that pave the way for future research and deserve being discussed. For instance, statistical analyses have highlighted specific problems in gaze direction modelling. This is a difficult hurdle to face. Some contextual rules (e.g., the prevalence of horizontal scanning) that have been advocated in the computer vision field (Torralba et al., 2006) and in the psychological literature (Tatler and Vincent, 2008), might fail in more ecological conditions, out of the lab and in dynamic environments. On the other hand, the ecology of animal movements is still struggling on the point (Viswanathan et al., 2011) in spite of an important body of research laid down over years. One solution could be that of a data-driven strategy (Le Meur and Coutrot, 2016; Hu et al., 2020), albeit raising in turn the problem of generalisability.

The model simulates fixation duration from first principles (Charnov's theorem) and achieves significant performance. Notwithstanding, it would be interesting to amend the lack of an explicit account for actual patch exploitation and within-patch item handling. One such example is facial expression processing of people engaged in the conversations. Expression perception is one fundamental mean for our understanding of and engagement in social interactions. This aspect is intimately related to the notion of value proposed in our work, which represents as a matter of fact a doorway to intertwine attention, cognition and emotion (Pessoa, 2008). Indeed, several studies have reported the influence of emotion on overt attention and emphasised the distinction between internally and externally located emotional cues; meanwhile, other studies have shown the reversed causal effect: attention can also affect emotional responses (Schomaker et al., 2017; Rubo and Gamer, 2018).

OFT is a general and appealing framework able to effectively detail the process of gaze deployment. However, this approach lacks the mechanistic description suitable to unravel the neural underpinnings of value-based decision making, for which dedicated models are available (Gold and Shadlen, 2007; Bogacz et al., 2006; Krajbich and Rangel, 2011; Noorani and Carpenter, 2016). Under such rationale, the very same problems outlined throughout this thesis can be recast into a so-called *Perceptual Decision Making* task, in which sensory information is gathered through the senses and then evaluated and integrated according to the current goals and internal state of the subject in order to take a decision. In this respect, the LATEST model proposed by Tatler et al. (2017) is worthy of mention; LATEST has its foundations in the LATER model (Carpenter, 1981; Noorani and Carpenter, 2016), which assumes that decision making is the result of the accumulation of evidence according to a linear function, until a threshold is reached and eventually an action is taken. LATEST simultaneously accounts for both when and where we look by evaluation of the relative benefit expected from moving the eyes to a new location compared with that expected by continuing to fixate the current target (Tatler et al., 2017). When the evidence in favor of a particular location of the

stimuli outweighs that of the current one, the eyes move to that location. As a result the duration of fixations is explicitly modeled as the latency of the decision making process i.e. the amount of time needed for one location's evidence to outweigh the others (saccade latency). From such point of view, the LATEST model shares some similarities with the principles advocated by MVT; indeed, it has been shown that under appropriate conditions, there might be a tight relationship between evidence accumulation models and foraging (Davidson and El Hady, 2019).

On the other hand, it is clear that a careful analysis of the model outlined in the present work may reveal some similarity with Reinforcement Learning (RL), albeit not adopting any explicit learning mechanism. A glimpse at Figure 4.2 unveils how it is possible to gain some insight on how the patch exploration/exploitation dilemma that the foraging eye has to solve, might be efficiently modelled through RL algorithms (Sutton et al., 1998). This link is further strengthened if considering the value/reward concepts that permeate OFT models. It is not surprising that recent research trends have shown how the behaviour of foraging animals can be well approximated by RL algorithms (Miller et al., 2017; Kolling and Akam, 2017). In particular the patch residence time (corresponding to fixations duration) may be learned and simulated effectively with such algorithms (Wawerla and Vaughan, 2009). In this respect, it would be interesting to recast the gaze deployment issue as a RL problem where the ultimate objective would be to learn to forage optimally on a (audio-)visual landscape. The main challenge here, would be to carefully design the reward function associated to each choice in order to accurately represent the task. Indeed, as discussed in Chapter 4, the goals of the observer may be endogenous and not easily definable explicitly. To this end, prior knowledge coming from foraging theory might be employed so to hand-craft a proper reward function. Yet, a more elegant solution could be provided by Inverse Reinforcement Learning (IRL) algorithms (Ng et al., 2000). In that case the main purpose is to learn an agent's objectives, values, or rewards functions from the demonstrations that could explain the expert's behavior. It's worth remarking that this approach would sensibly deviate from the one adopted here, as the current implementation does not rely on any learning procedure for the patch value estimation, but infers it by assuming that the value is proportional to the amount of "real" fixations that fall in a given patch at any given time. In this sense an IRL approach could be an interesting alternative.

The proposed model has been presented and tested on the particular task of attention allocation on social video clips. Despite being a widely spread kind of task to deal with and representing a crucial aspect in social robotics (Admoni and Scassellati, 2017), some might see this as a big limitation. A quantitative account on the possibility to extend the proposed model to a wider class of stimuli, tasks and conditions will be part of future works. However, the general principles at the foundation of the adopted method suggest that this might be successful in explaining gaze behaviour even in more general cases. These facts need to be further investigated.

To sum up, despite such limitations, the results presented in this study allow to draw at least two general conclusions.

The first lies in that when we engage with the computational modelling of attention in multimodal scenarios with rich semantics, we should not overstate the role of classic salience. Concentrating all research efforts by mostly focusing on subtle improvements of such techniques (whose statistical significance is at best questionable), under the

wishful assumption that these will be predictive of actual gaze allocation, might not be the optimal strategy.

The second and artful lesson to learn, is that general models of gaze deployment are appealing, indeed elegant and explainable. Nevertheless, they should be general as to the foundational principles and rationales, albeit not generic. Caution suggests that the many dimensions of gaze dynamics are to be specifically accounted for, if the similarity to human gaze behaviour is the ultimate goal.

Bibliography

Admoni, H. and Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63.

Aloimonos, J., Weiss, I., and Bandyopadhyay, A. (1988). Active vision. *International journal of computer vision*, 1(4):333–356.

Anderson, B. A. (2013). A value-driven mechanism of attentional selection. *Journal of vision*, 13(3).

Anderson, N. C., Anderson, F., Kingstone, A., and Bischof, W. F. (2015). A comparison of scanpath comparison methods. *Behavior research methods*, 47(4):1377–1392.

Anderson, N. C., Bischof, W. F., Laidlaw, K. E., Risko, E. F., and Kingstone, A. (2013). Recurrence quantification analysis of eye movements. *Behavior research methods*, 45(3):842–856.

Assens, M., Giro-i Nieto, X., McGuinness, K., and O'Connor, N. E. (2018a). Pathgan: visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0.

Assens, M., Giro-i Nieto, X., McGuinness, K., and O'Connor, N. E. (2018b). Scanpath and saliency prediction on 360 degree images. *Signal Processing: Image Communication*, 69:8–14.

Averbeck, B. B. (2015). Theory of choice in bandit, information sampling and foraging tasks. *PLoS computational biology*, 11(3).

Awh, E., Belopolsky, A. V., and Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in cognitive sciences*, 16(8):437–443.

Ballard, D. (1991). Animate vision. *Artificial intelligence*, 48(1):57–86.

Bao, W. and Chen, Z. (2020a). Human scanpath prediction based on deep convolutional saccadic model. *Neurocomputing*.

Bao, W. and Chen, Z. (2020b). Human scanpath prediction based on deep convolutional saccadic model. *Neurocomputing*, 404:154 – 164.

Bargary, G., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Hogg, R. E., and Mollon, J. D. (2017). Individual differences in human eye movements: An oculomotor signature? *Vision research*, 141:157–169.

Barthelmé, S., Trukenbrod, H., Engbert, R., and Wichmann, F. (2013). Modeling fixation locations using spatial point processes. *Journal of vision*, 13(12):1–1.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282.

Bartumeus, F. and Catalan, J. (2009). Optimal search behavior and classic foraging theory. *Journal of Physics A: Mathematical and Theoretical*, 42:434002.

Benhamou, S. (2007). How many animals really do the Lévy walk? *Ecology*, 88(8):1962–1969.

Bénichou, O., Loverdo, C., Moreau, M., and Voituriez, R. (2006). Two-dimensional intermittent search processes: An alternative to lévy flight strategies. *Physical Review E*, 74(2):020102.

Berridge, K. C. and Robinson, T. E. (2003). Parsing reward. *Trends in neurosciences*, 26(9):507–513.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654.

Boccignone, G. (2008). Nonparametric bayesian attentive video analysis. In *Proc. 19th International Conference on Pattern Recognition, ICPR 2008*, pages 1–4. IEEE Press.

Boccignone, G. (2016). A probabilistic tour of visual attention and gaze shift computational models. In *Int. Workshop Vision over vision: man, monkey, machines, and network models*. Osaka, Japan.

Boccignone, G., Cuculo, V., and D'Amelio, A. (2019a). Problems with saliency maps. In Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., and Sebe, N., editors, *Image Analysis and Processing – ICIAP 2019*, pages 35–46, Cham. Springer International Publishing.

Boccignone, G., Cuculo, V., D'Amelio, A., Grossi, G., and Lanzarotti, R. (2019b). Give ear to my face: Modelling multimodal attention to social interactions. In Leal-Taixé, L. and Roth, S., editors, *Computer Vision – ECCV 2018 Workshops*, pages 331–345. Springer International Publishing, Cham.

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

Boccignone, G. and Ferraro, M. (2014). Ecological sampling of gaze shifts. *IEEE Trans. on Cybernetics*, 44(2):266–279.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, 113(4):700.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Borji, A. (2019). Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*.

Borji, A. and Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207.

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

Borji, A. and Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*. arXiv preprint arXiv:1505.03581.

Borji, A., Sihite, D. N., and Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *Image Processing, IEEE Transactions on*, 22(1):55–69.

Brandt, S. A. and Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, 9(1):27–38.

Breed, G. A., Golson, E. A., and Tinker, M. T. (2017). Predicting animal home-range structure and transitions using a multistate ornstein-uhlenbeck biased random walk. *Ecology*, 98(1):32–47.

Brillinger, D. R., Preisler, H. K., Ager, A. A., Kie, J. G., and Stewart, B. S. (2002). Employing stochastic differential equations to model wildlife motion. *Bulletin of the Brazilian Mathematical Society*, 33(3):385–408.

Brockmann, D. and Geisel, T. (2000). The ecology of gaze shifts. *Neurocomputing*, 32:643–650.

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462–465.

Bruce, N. D., Wloka, C., Frosst, N., Rahman, S., and Tsotsos, J. K. (2015). On computational modeling of visual saliency: Examining what's right, and what's left. *Vision research*, 116:95–112.

Butko, N. J. and Movellan, J. R. (2008). I-pomdp: An infomax model of eye movement. In *2008 7th IEEE International Conference on Development and Learning*, pages 139–144. IEEE.

Bylinskii, Z., DeGennaro, E., Rajalingham, R., Ruda, H., Zhang, J., and Tsotsos, J. (2015). Towards the quantitative evaluation of visual attention models. *Vision research*, 116:258–268.

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., and Durand, F. (2019). What do different evaluation metrics tell us about saliency models? *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 41(3):740–757.

Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., and Durand, F. (2016). Where should saliency models look next? In *European Conference on Computer Vision*, pages 809–824. Springer.

Cain, M. S., Vul, E., Clark, K., and Mitroff, S. R. (2012). A bayesian optimal foraging model of human visual search. *Psychological science*, 23(9):1047–1054.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D., and David, A. S. (1997). Activation of auditory cortex during silent lipreading. *science*, 276(5312):593–596.

Canosa, R. (2009). Real-world vision: Selective perception and task. *ACM Transactions on Applied Perception*, 6(2):11.

Carpenter, R. H. (1981). Oculomotor procrastination. *movement: Cognition and Visual Perception*.

Cerf, M., Frady, E., and Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, 9(12).

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

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical population biology*, 9(2):129–136.

Chelazzi, L., Eštočinová, J., Calletti, R., Gerfo, E. L., Sani, I., Della Libera, C., and Santandrea, E. (2014). Altering spatial priority maps via reward-based learning. *Journal of Neuroscience*, 34(25):8594–8604.

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

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979.

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

Chung, J. S. and Zisserman, A. (2016). Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*.

Chung, J. S. and Zisserman, A. (2017). Lip reading in profile. BMVC.

Clavelli, A., Karatzas, D., Lladós, J., Ferraro, M., and Boccignone, G. (2014). Modelling task-dependent eye guidance to objects in pictures. *Cognitive Computation*, 6(3):558–584.

Cliff, N. (2014). *Ordinal methods for behavioral data analysis*. Psychology Press.

Codling, E. A., Plank, M. J., and Benhamou, S. (2008). Random walk models in biology. *Journal of the Royal Society Interface*, 5(25):813–834.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Corbetta, M. (1998). Frontoparietal cortical networks for directing attention and the eye to visual locations: identical, independent, or overlapping neural systems? *Proceedings of the National Academy of Sciences*, 95(3):831–838.

Coutrot, A. and Guyader, N. (2014a). An audiovisual attention model for natural conversation scenes. In *Proc. IEEE International Conference on Image processing (ICIP)*, pages 1100–1104. IEEE.

Coutrot, A. and Guyader, N. (2014b). How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of vision*, 14(8):5–5.

Coutrot, A. and Guyader, N. (2015). An efficient audiovisual saliency model to predict eye positions when looking at conversations. In *23rd European Signal Processing Conference*, pages 1531–1535.

Coutrot, A. and Guyader, N. (2016). Multimodal saliency models for videos. In *From Human Attention to Computational Attention*, pages 291–304. Springer.

Cristino, F., Mathôt, S., Theeuwes, J., and Gilchrist, I. D. (2010). Scanmatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, 42(3):692–700.

Cuculo, V., D'Amelio, A., Lanzarotti, R., and Boccignone, G. (2018). Personality gaze patterns unveiled via automatic relevance determination. In *Federation of International Conferences on Software Technologies: Applications and Foundations*, pages 171–184. Springer.

Davidson, J. D. and El Hady, A. (2019). Foraging as an evidence accumulation process. *PLoS computational biology*, 15(7):e1007060.

de Croon, G., Sprinkhuizen-Kuyper, I. G., and Postma, E. O. (2009). Comparing active vision models. *Image and Vision Computing*, 27(4):374–384.

de Knegt, H., Hengeveld, G., van Langevelde, F., de Boer, W., and Kirkman, K. (2007). Patch density determines movement patterns and foraging efficiency of large herbivores. *Behavioral Ecology*, 18(6):1065–1072.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

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

Deubel, H., Schneider, W. X., et al. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision research*, 36(12):1827–1838.

Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., and Holmqvist, K. (2012). It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44(4):1079–1100.

Donnarumma, F., Costantini, M., Ambrosini, E., Friston, K., and Pezzulo, G. (2017). Action perception as hypothesis testing. *Cortex*, 89:45–60.

Dorr, M., Bohme, M., Martinetz, T., Gegenfurtner, K., and Barth, E. (2005). Variability of eye movements on natural videos. In *Proc. 8th Tubingen Perception Conference*, page 162, Tubingen, Germany.

Dorr, M., Martinetz, T., Gegenfurtner, K., and Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(10).

Edwards, A. M., Phillips, R. A., Watkins, N. W., Freeman, M. P., Murphy, E. J., Afanasyev, V., Buldyrev, S. V., da Luz, M. G., Raposo, E. P., Stanley, H. E., et al. (2007). Revisiting lévy flight search patterns of wandering albatrosses, bumblebees and deer. *Nature*, 449(7165):1044–1048.

Egeth, H. E. and Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual review of psychology*, 48(1):269–297.

Ehinger, K. A. and Wolfe, J. M. (2016). When is it time to move to the next map? optimal foraging in guided visual search. *Attention, Perception, & Psychophysics*, 78(7):2135–2151.

Einhäuser, W., Spain, M., and Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, 8(14).

Einstein, A. (1905). On the motion required by the molecular kinetic theory of heat of small particles suspended in a stationary liquid. *Annalen der Physik*, 17:549–560.

Einstein, A. (1906). Zur theorie der brownschen bewegung. *Annalen der Physik*, 324(2):371–381.

Elazary, L. and Itti, L. (2010). A bayesian model for efficient visual search and recognition. *Vision research*, 50(14):1338–1352.

Emlen, J. M. (1966). The role of time and energy in food preference. *The American Naturalist*, 100(916):611–617.

Engbert, R. (2006). Microsaccades: A microcosm for research on oculomotor control, attention, and visual perception. *Progress in brain research*, 154:177–192.

Engbert, R. and Kliegl, R. (2004). Microsaccades keep the eyes' balance during fixation. *Psychological science*, 15(6):431–431.

Engbert, R., Mergenthaler, K., Sinn, P., and Pikovsky, A. (2011). An integrated model of fixational eye movements and microsaccades. *Proceedings of the National Academy of Sciences*, 108(39):E765–E770.

Evangelopoulos, G., Rapantzikos, K., Maragos, P., Avrithis, Y., and Potamianos, A. (2008). Audiovisual attention modeling and salient event detection. In *Multimodal Processing and Interaction*, pages 1–21. Springer.

Fecteau, J. H. and Munoz, D. P. (2006). Salience, relevance, and firing: a priority map for target selection. *Trends in cognitive sciences*, 10(8):382–390.

Findlay, J. M. and Walker, R. (1999). A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences*, 22(4):661–674.

Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., and Kingstone, A. (2010). Gaze allocation in a dynamic situation: Effects of social status and speaking. *Cognition*, 117(3):319–331.

Foulsham, T. and Underwood, G. (2008). What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, 8(2).

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

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

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1-3):70–87.

Furnari, A., Farinella, G. M., and Battiato, S. (2014). An experimental analysis of saliency detection with respect to three saliency levels. In *European Conference on Computer Vision*, pages 806–821. Springer.

Galton, F. (1894). *Natural inheritance*. Macmillan and Company.

Gardiner, C. (2011). *Handbook of stochastic methods*. Springer–Verlag, Berlin, Germany.

Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., and He, K. (2018). Detectron. `https://github.com/facebookresearch/detectron`.

Gnedenko, B. and Kolmogórov, A. (1954). *Limit distributions for sums of independent random variables*. Addison-Wesley Pub. Co.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annual review of neuroscience*, 30.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *nature*, 453(7196):779–782.

Green, R. F. (1980). Bayesian birds: a simple example of oaten's stochastic model of optimal foraging. *Theoretical Population Biology*, 18(2):244–256.

Groen, I. I., Silson, E. H., and Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Phil. Trans. R. Soc. B*, 372(1714):20160102.

Grossman, R. B., Mertens, J., and Zane, E. (2019). Perceptions of self and other: Social judgments and gaze patterns to videos of adolescents with and without autism spectrum disorder. *Autism*, 23(4):846–857.

Guy, N., Azulay, H., Kardosh, R., Weiss, Y., Hassin, R. R., Israel, S., and Pertzov, Y. (2019). A novel perceptual trait: Gaze predilection for faces during visual exploration. *Scientific reports*, 9(1):1–12.

Han, P., Saunders, D. R., Woods, R. L., and Luo, G. (2013). Trajectory prediction of saccadic eye movements using a compressed exponential model. *Journal of vision*, 13(8):27–27.

Harel, J., Koch, C., and Perona, P. (2007). Graph-based visual saliency. In *Advances in neural information processing systems*, volume 19, pages 545–552, Cambridge, MA. MIT Press.

Harris, K. J. and Blackwell, P. G. (2013). Flexible continuous-time modelling for heterogeneous animal movement. *Ecological Modelling*, 255:29–37.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Heinke, D. and Backhaus, A. (2011). 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.

Heinke, D. and Humphreys, G. W. (2003). Attention, spatial representation, and visual neglect: simulating emergent attention and spatial memory in the selective attention for identification model (saim). *Psychological review*, 110(1):29.

Henderson, J. M. (2017). Gaze control as prediction. *Trends in cognitive sciences*, 21(1):15–23.

Henderson, J. M. and Luke, S. G. (2014). Stable individual differences in saccadic eye movements during reading, pseudoreading, scene viewing, and scene search. *Journal of Experimental Psychology: Human Perception and Performance*, 40(4):1390.

Hess, M. R. and Kromrey, J. D. (2004). Robust confidence intervals for effect sizes: A comparative study of cohen'sd and cliff's delta under non-normality and heterogeneous variances. In *annual meeting of the American Educational Research Association*, pages 1–30.

Higham, D. (2001). An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM review*, pages 525–546.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science*, 30(1):3–41.

Hoffman, J. E. (1998). Visual attention and eye movements. *Attention*, 31:119–153.

Hu, P. and Ramanan, D. (2017). Finding tiny faces. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1522–1530. IEEE.

Hu, Z., Li, S., Zhang, C., Yi, K., Wang, G., and Manocha, D. (2020). Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1902–1911.

Humphreys, G. W. and Muller, H. J. (1993). Search via recursive rejection (serr): A connectionist model of visual search. *Cognitive Psychology*, 25(1):43–110.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews - Neuroscience*, 2:1–11.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1254–1259.

Iwasa, Y., Higashi, M., and Yamamura, N. (1981). Prey distribution as a factor determining the choice of optimal foraging strategy. *The American Naturalist*, 117(5):710–723.

James, W. (1890). *The Principles of Psychology*. Dover Publications.

Jarodzka, H., Holmqvist, K., and Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)*, ETRA '10, pages 211–218, New York, NY, USA. ACM.

Jiang, L., Xu, M., Ye, Z., and Wang, Z. (2015a). Image saliency detection with sparse representation of learnt texture atoms. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 54–62.

Jiang, M., Huang, S., Duan, J., and Zhao, Q. (2015b). Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jording, M., Engemann, D., Eckert, H., Bente, G., and Vogeley, K. (2019). Distinguishing social from private intentions through the passive observation of gaze cues. *Frontiers in Human Neuroscience*, 13:442.

Judd, T., Durand, F., and Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*.

Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *IEEE 12th International conference on Computer Vision*, pages 2106–2113. IEEE.

Kaya, E. M. and Elhilali, M. (2017). Modelling auditory attention. *Phil. Trans. R. Soc. B*, 372(1714):20160101.

Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology*, 15(21):1943–1947.

Kazimierski, L. D., Abramson, G., and Kuperman, M. N. (2016). The movement of a forager: strategies for the efficient use of resources. *The European Physical Journal B*, 89(10):232.

Kienzle, W., Wichmann, F. A., Franz, M. O., and Schölkopf, B. (2006). A nonparametric approach to bottom-up visual saliency. In *Advances in neural information processing systems*, pages 689–696.

Klink, P. C., Jentgens, P., and Lorteije, J. A. (2014). Priority maps explain the roles of value, attention, and salience in goal-oriented behavior. *Journal of Neuroscience*, 34(42):13867–13869.

Kloeden, P. E. and Platen, E. (2013). *Numerical solution of stochastic differential equations*, volume 23. Springer Science & Business Media.

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

Kolling, N. and Akam, T. (2017). (reinforcement?) learning to forage optimally. *Current opinion in neurobiology*, 46:162–169.

Kondo, H. M., van Loon, A. M., Kawahara, J.-I., and Moore, B. C. J. (2017). Auditory and visual scene analysis: an overview. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714).

Kong, P., Mancas, M., Thuon, N., Kheang, S., and Gosselin, B. (2018). Do deep-learning saliency models really model saliency? In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2331–2335. IEEE.

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

Krajbich, I. and Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33):13852–13857.

Kriegstein, K. v., Kleinschmidt, A., Sterzer, P., and Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of cognitive neuroscience*, 17(3):367–376.

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

Kümmerer, M., Theis, L., and Bethge, M. (2014). Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*.

Kümmerer, M., Wallis, T. S., and Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059.

Kummerer, M., Wallis, T. S., Gatys, L. A., and Bethge, M. (2017). Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4789–4798.

Kümmerer, M., Wallis, T. S. A., and Bethge, M. (2018). Saliency benchmarking made easy: Separating models, maps and metrics. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 798–814. Springer International Publishing.

Kunert, J., Montag, A., and Pöhlmann, S. (2001). The quincunx: history and mathematics. *Statistical Papers*, 42(2):143–169.

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, 25(3):296 – 324.

Lang, C., Liu, G., Yu, J., and Yan, S. (2011). Saliency detection by multitask sparsity pursuit. *IEEE transactions on image processing*, 21(3):1327–1338.

Lang, C., Liu, G., Yu, J., and Yan, S. (2012). Saliency detection by multitask sparsity pursuit. *IEEE Transactions on Image Processing*, 21(3):1327–1338.

Le Meur, O. and Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266.

Le Meur, O. and Coutrot, A. (2016). Introducing context-dependent and spatially-variant viewing biases in saccadic models. *Vision Research*, 121:72–84.

Le Meur, O. and Liu, Z. (2015). Saccadic model of eye movements for free-viewing condition. *Vision research*, 116:152–164.

Lee, K., Ognibene, D., Chang, H. J., Kim, T.-K., and Demiris, Y. (2015). Stare: Spatio-temporal attention relocation for multiple structured activities detection. *IEEE Transactions on Image Processing*, 24(12):5916–5927.

Lemons, D. S. (2002). *An introduction to stochastic processes in physics*. JHU Press.

Levene, H. (1961). Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, pages 279–292.

Li, X., Wang, W., Hou, W., Liu, R.-Z., Lu, T., and Yang, J. (2018). Shape robust text detection with progressive scale expansion network. *arXiv preprint arXiv:1806.02559*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

MacArthur, R. H. and Pianka, E. R. (1966). On optimal use of a patchy environment. *The American Naturalist*, 100(916):603–609.

Mackay, M., Cerf, M., and Koch, C. (2012). Evidence for two distinct mechanisms directing gaze in natural scenes. *Journal of Vision*, 12(4):9–9.

Makarava, N., Bettenbühl, M., Engbert, R., and Holschneider, M. (2012). Bayesian estimation of the scaling parameter of fixational eye movements. *EPL*, 100(4):40003.

Mandelbrot, B. (1960). The pareto-levy law and the distribution of income. *International Economic Review*, 1(2):79–106.

Mandelbrot, B. (1963). The variation of certain speculative prices. *The journal of business*, 36(4):394–419.

Mandelbrot, B. B. and Van Ness, J. W. (1968). Fractional brownian motions, fractional noises and applications. *SIAM review*, 10(4):422–437.

Marat, S., Rahman, A., Pellerin, D., Guyader, N., and Houzet, D. (2013). Improving visual saliency by adding Ôface feature mapÕand Ôcenter biasÕ. *Cognitive Computation*, 5(1):63–75.

Marlow, C. A., Viskontas, I. V., Matlin, A., Boydston, C., Boxer, A., and Taylor, R. P. (2015). Temporal structure of human gaze dynamics is invariant during free viewing. *PloS one*, 10(9):e0139379.

marquis de Laplace, P. S. (1902). *A philosophical essay on probabilities*. Wiley.

Martinez-Conde, S., Otero-Millan, J., and Macknik, S. L. (2013). The impact of microsaccades on vision: towards a unified theory of saccadic function. *Nature Reviews Neuroscience*, 14(2):83–96.

Mathe, S. and Sminchisescu, C. (2015). Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(7):1408–1424.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746.

McNamara, J. (1982). Optimal patch use in a stochastic environment. *Theoretical Population Biology*, 21(2):269–288.

McNamara, J. and Houston, A. (1985). A simple model of information use in the exploitation of patchily distributed food. *Animal Behaviour*, 33(2):553–560.

Méndez, V., Campos, D., and Bartumeus, F. (2014). *Stochastic Foundations in Movement Ecology: Anomalous Diffusion, Front Propagation and Random Searches*. Springer Series in Synergetics. Springer-Verlag, Berlin, Heidelberg.

Miller, M. L., Ringelman, K. M., Eadie, J. M., and Schank, J. C. (2017). Time to fly: A comparison of marginal value theorem approximations in an agent-based model of foraging waterfowl. *Ecological Modelling*, 351:77–86.

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Frontiers in computational neuroscience*, 10:56.

Mozer, M. C. (1987). *Early parallel processing in reading: A connectionist approach.* Lawrence Erlbaum Associates, Inc.

Najemnik, J. and Geisler, W. (2005). Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391.

Napoletano, P., Boccignone, G., and Tisato, F. (2015). Attentive monitoring of multiple video streams driven by a bayesian foraging strategy. *IEEE Trans. on Image Processing*, 24(11):3266 – 3281.

Nassauer, A. and Legewie, N. M. (2019). Analyzing 21st century video data on situational dynamics—issues and challenges in video data analysis. *Social Sciences*, 8(3):100.

Nelson, E. (1967). *Dynamical theories of Brownian motion*. Princeton University Press, Princeton, NJ.

Nemenyi, P. B. (1963). Distribution-free multiple comparisons. Phd thesis, Princeton University.

Ng, A. Y., Russell, S. J., et al. (2000). Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2.

Nguyen, T. V., Zhao, Q., and Yan, S. (2018). Attentive systems: A survey. *International Journal of Computer Vision*, 126(1):86–110.

Noorani, I. and Carpenter, R. (2016). The later model of reaction time and decision. *Neuroscience & Biobehavioral Reviews*, 64:229–251.

Nuthmann, A., Smith, T. J., Engbert, R., and Henderson, J. M. (2010). Crisp: a computational model of fixation durations in scene viewing. *Psychological review*, 117(2):382.

Ognibene, D. and Baldassare, G. (2014). Ecological active vision: four bioinspired principles to integrate bottom–up and adaptive top–down attention tested with a simple camera-arm robot. *IEEE transactions on autonomous mental development*, 7(1):3–25.

Ognibene, D., Chinellato, E., Sarabia, M., and Demiris, Y. (2013). Contextual action recognition and target localization with an active allocation of attention on a humanoid robot. *Bioinspiration & biomimetics*, 8(3):035002.

Ognibene, D. and Demiris, Y. (2013). Towards active event recognition. In *IJCAI*, volume 13, pages 2495–501.

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

Onat, S., Libertus, K., and König, P. (2007). Integrating audiovisual information for the control of overt attention. *Journal of Vision*, 7(10):11–11.

Otero-Millan, J., Macknik, S. L., Langston, R. E., and Martinez-Conde, S. (2013). An oculomotor continuum from exploration to fixation. *Proceedings of the National Academy of Sciences*, 110(15):6175–6180.

Oud, J. H. and Singer, H. (2008). Continuous time modeling of panel data: Sem versus filter techniques. *Statistica Neerlandica*, 62(1):4–28.

Over, E., Hooge, I., Vlaskamp, B., and Erkelens, C. (2007). Coarse-to-fine eye movement strategy in visual search. *Vision Research*, 47:2272–2280.

Paletta, L., Fritz, G., and Seifert, C. (2005). Q-learning of sequential attention for visual object recognition from informative local descriptors. In *Proceedings of the 22nd international conference on Machine learning*, pages 649–656.

Parr, T. and Friston, K. J. (2018). Active inference and the anatomy of oculomotion. *Neuropsychologia*, 111:334–343.

Pekkanen, J. and Lappi, O. (2017). A new and general approach to signal denoising and eye movement classification based on segmented linear regression. *Scientific reports*, 7(1):1–13.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, 9(2):148–158.

Peters, R. J., Iyer, A., Itti, L., and Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416.

Phaf, R. H., Van der Heijden, A., and Hudson, P. T. (1990). Slam: A connectionist model for attention in visual selection tasks. *Cognitive Psychology*, 22(3):273–341.

Pires, K. and Simon, G. (2015). Youtube live and twitch: a tour of user-generated live streaming systems. In *Proceedings of the 6th ACM multimedia systems conference*, pages 225–230.

Richardson, L. F. (1926). Atmospheric diffusion shown on a distance-neighbour graph. *Proceedings of the Royal Society of London. Series A*, 110(756):709–737.

Riche, N., Duvinage, M., Mancas, M., Gosselin, B., and Dutoit, T. (2013). Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE international conference on computer vision*, pages 1153–1160.

Rodríguez-Gironés, M. A. and Vasquez, R. A. (1997). Density-dependent patch exploitation and acquisition of environmental information. *Theoretical Population Biology*, 52(1):32–42.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (2007). Do you see what i am saying? exploring visual enhancement of speech comprehension in noisy environments. *Cerebral cortex*, 17(5):1147–1153.

Rothenstein, A. L. and Tsotsos, J. K. (2008). Attention links sensing to recognition. *Image and Vision Computing*, 26(1):114–126.

Rothkegel, L. O., Trukenbrod, H. A., Schütt, H. H., Wichmann, F. A., and Engbert, R. (2017). Temporal evolution of the central fixation bias in scene viewing. *Journal of vision*, 17(13):3–3.

Rothkopf, C., Ballard, D., and Hayhoe, M. (2007). Task and context determine where you look. *Journal of Vision*, 7(14).

Rubo, M. and Gamer, M. (2018). Social content and emotional valence modulate gaze fixations in dynamic scenes. *Scientific reports*, 8(1):1–11.

Rukavicka, J. (2014). Rejection of laplace's demon. *The American Mathematical Monthly*, 121(6):498–498.

Schomaker, J., Walper, D., Wittmann, B. C., and Einhäuser, W. (2017). Attention in natural scenes: affective-motivational factors guide gaze independently of visual salience. *Vision Research*, 133:161–175.

Schütt, H. H., Rothkegel, L. O., Trukenbrod, H. A., Engbert, R., and Wichmann, F. A. (2019). Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. *Journal of vision*, 19(3):1–1.

Schütz, A., Braun, D., and Gegenfurtner, K. (2011). Eye movements and perception: A selective review. *Journal of Vision*, 11(5).

Seo, H. and Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):1–27.

Serences, J. T. and Yantis, S. (2006). Selective visual attention and perceptual coherence. *Trends in cognitive sciences*, 10(1):38–45.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Shepherd, S. V. and Platt, M. L. (2007). Spontaneous social orienting and gaze following in ringtailed lemurs (lemur catta). *Animal Cognition*, 11(1):13.

Shic, F., Scassellati, B., Lin, D., and Chawarska, K. (2007). Measuring context: The gaze patterns of children with autism evaluated from the bottom-up. In *2007 IEEE 6th International Conference on Development and Learning*, pages 70–75. IEEE.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in cognitive sciences*, 12(5):182–186.

Simons, D. J. and Levin, D. T. (1997). Change blindness. *Trends in cognitive sciences*, 1(7):261–267.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Staab, J. P. (2014). The influence of anxiety on ocular motor control and gaze. *Current opinion in neurology*, 27(1):118–124.

Stephens, D. W. (1986). *Foraging theory*. Princeton University Press.

Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215.

Sun, W., Chen, Z., and Wu, F. (2019). Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sun, X., Yao, H., Ji, R., and Liu, X.-M. (2014). Toward statistical modeling of saccadic eye-movement and visual saliency. *IEEE Transactions on Image Processing*, 23(11):4649–4662.

Sutton, R. S., Barto, A. G., et al. (1998). *Introduction to reinforcement learning*, volume 135. MIT press Cambridge.

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

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

Tatler, B. and Vincent, B. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, 2(2):1–18.

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

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

Tatler, B. W., Brockmole, J. R., and Carpenter, R. H. (2017). Latest: A model of saccadic decisions in space and time. *Psychological Review*, 124(3):267.

Tavakoli, H. R., Borji, A., Rahtu, E., and Kannala, J. (2019). DAVE: A deep audio-visual embedding for dynamic saliency prediction. *CoRR*, abs/1905.10693.

Torralba, A. (2003). Contextual priming for object detection. *Int. J. of Comp. Vis.*, 53:153–167.

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

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1373):1295–1306.

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

Truong, A. and Agrawala, M. (2019). A tool for navigating and editing 360 video of social conversations into shareable highlights. In *Proceedings of the 45th Graphics Interface Conference on Proceedings of Graphics Interface 2019*, pages 1–9. Canadian Human-Computer Communications Society.

Tseng, P.-H., Carmi, R., Cameron, I. G., Munoz, D. P., and Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of vision*, 9(7):4–4.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1):507–545.

Uhlenbeck, G. E. and Ornstein, L. S. (1930). On the theory of the brownian motion. *Physical review*, 36(5):823.

Unema, P., Pannasch, S., Joos, M., and Velichkovsky, B. (2007). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition*, 12(3):473–494.

van Beers, R. (2007). The sources of variability in saccadic eye movements. *The Journal of Neuroscience*, 27(33):8757–8770.

Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., and Theeuwes, J. (2008). Audiovisual events capture attention: Evidence from temporal order judgments. *Journal of vision*, 8(5):2–2.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of financial economics*, 5(2):177–188.

Vig, E., Dorr, M., and Cox, D. (2014). Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2798–2805.

Viola, P. and Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Viswanathan, G., Afanasyev, V., Buldyrev, S., Murphy, E., Prince, P., and Stanley, H. (1996). Lévy flight search patterns of wandering albatrosses. *Nature*, 381(6581):413–415.

Viswanathan, G., Raposo, E., and Da Luz, M. (2008). Lévy flights and superdiffusion in the context of biological encounters and random searches. *Physics of Life Reviews*, 5(3):133–150.

Viswanathan, G. M., Buldyrev, S. V., Havlin, S., Da Luz, M., Raposo, E., and Stanley, H. E. (1999). Optimizing the success of random searches. *nature*, 401(6756):911–914.

Viswanathan, G. M., Da Luz, M. G., Raposo, E. P., and Stanley, H. E. (2011). *The physics of foraging: an introduction to random searches and biological encounters*. Cambridge University Press, Cambridge, UK.

Von Helmholtz, H. (1867). Treatise on physiological optics vol. iii.

Von Smoluchowski, M. (1906). Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen. *Annalen der physik*, 326(14):756–780.

Wang, W., Chen, C., Wang, Y., Jiang, T., Fang, F., and Yao, Y. (2011). Simulating human saccadic scanpaths on natural images. In *CVPR 2011*, pages 441–448. IEEE.

Wawerla, J. and Vaughan, R. T. (2009). Robot task switching under diminishing returns. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5033–5038. IEEE.

Wiener, N. (1930). Generalized harmonic analysis. *Acta mathematica*, 55(1):117–258.

Wilkinson, N., Paikan, A., Gredebäck, G., Rea, F., and Metta, G. (2014). Staring us in the face? an embodied theory of innate face preference. *Developmental Science*, 17(6):809–825.

Wischnewski, M., Belardinelli, A., Schneider, W., and Steil, J. (2010). Where to Look Next? Combining Static and Dynamic Proto-objects in a TVA-based Model of Visual Attention. *Cognitive Computation*, 2(4):326–343.

Wloka, C., Kotseruba, I., and Tsotsos, J. K. (2018). Active fixation control to predict saccade sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3184–3193.

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

Wolfe, J. M. (2013). When is it time to move to the next raspberry bush? Foraging rules in human visual search. *Journal of vision*, 13(3):10.

Wosniack, M. E., Santos, M. C., Raposo, E. P., Viswanathan, G. M., and da Luz, M. G. (2017). The evolutionary origins of lévy walk foraging. *PLoS computational biology*, 13(10):e1005774.

Xia, C., Han, J., Qi, F., and Shi, G. (2019). Predicting human saccadic scanpaths based on iterative representation learning. *IEEE Transactions on Image Processing*, pages 1–1.

Xia, C. and Quan, R. (2020). Predicting saccadic eye movements in free viewing of webpages. *IEEE Access*, 8:15598–15610.

Xu, M., Liu, Y., Hu, R., and He, F. (2018). Find who to look at: Turning from action to saliency. *IEEE Transactions on Image Processing*, 27(9):4529–4544.

Yan, J., Zhu, M., Liu, H., and Liu, Y. (2010). Visual saliency detection via sparsity pursuit. *Signal Processing Letters, IEEE*, 17(8):739–742.

Yang, S. C.-H., Wolpert, D. M., and Lengyel, M. (2016). Theoretical perspectives on active sensing. *Current Opinion in Behavioral Sciences*, 11:100–108.

Yarbus, A. (1967). *Eye Movements and Vision*. Plenum Press, New York.

Yu, J.-G., Zhao, J., Tian, J., and Tan, Y. (2014). Maximal entropy random walk for region-based visual saliency. *IEEE Transactions on Cybernetics*, 44(9):1661–1672.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zanca, D., Melacci, S., and Gori, M. (2019). Gravitational laws of focus of attention. *IEEE transactions on pattern analysis and machine intelligence*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.