UNIVERSITÀ
DEGLI STUDI
FIRENZE

PhD Program in Smart Computing
Dipartimento di Ingegneria dell'Informazione (DINFO)

Towards Laws of Visual Attention

Dario Zanca

Dissertation presented in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Smart Computing

PhD Program in Smart Computing
University of Florence, University of Pisa, University of Siena

Towards Laws of Visual Attention

Dario Zanca

Advisor:

Prof. Marco Gori

Head of the PhD Program:

Prof. Paolo Frasconi

Evaluation Committee:
Prof. Claudio M. Privitera, *University of Berkely*
Prof. Alessandro Villa, *Université de Lausanne*

XXXI ciclo — October 2019

To my family

Acknowledgement

I like to say that I ended up studying artificial intelligence by accident. I was fascinated by the reading of the seminal works of Turing and Searle at the time of my master's degree. I knew nothing more about it. Three years later, I think it was one of the most exciting adventures of my entire career. I still don't know much about it, but it is enough to say that it changed the way I look at the world and my own life. Investigating human intelligence in the hope of identifying its true computational components, understanding its ability to develop situated knowledge, generalize and adapt fast to new different environments, that proceeds through perception, interaction and reasoning - is among the noblest of the missions every human being should devote themselves to. This makes me proud to have expanded our knowledge in this area, even if by an infinitesimal.

But no one can do anything on their own. There were so many great people that helped along the way. I would like to acknowledge and to send thanks to a few of them.

I would like to thank first of all my supervisor Marco Gori, for guiding me as a researcher and for providing me with so many inspiring ideas during this time. Besides my advisor, I would like to thank the rest of my thesis committee: professors Roberto Serra and Marcello Pelillo for their insightful comments and encouragement, but also for the hard questions which incented me to widen my research from various perspectives; professors Alessandro Villa and Claudio Privitera for their careful revision that has brought value to my work.

My sincere thanks to Dr. Alessandra Rufa and her team for the countless discussions on eye movements, attention and human perception, and also for giving me access to their laboratory and collect data from human subjects. Thanks to the Caltech Vision Lab, who provided me research facilities during my period in Pasadena and taught me a lot about computer vision. Thanks to Stefano Melacci for helping me make the online demo of Eymol. Thanks to Mattia Bongini, who helped me when I was in trouble with some variational problems.

My time at the Siena Artificial Intelligence Lab was great because of the great colleagues. Alessandro Rossi, somehow my mentor, who was a navigated PhD when I first came here. Giuseppe Marra with whom I shared many scientific fantasies but above all the passion for homemade food. Andrea Zugarini, who shared with

me two exciting hackathons: we know the secrets behind the victories. Vincenzo Laveglia who sat at his desk in front of me for three years: a certainty. Francesco Giannini that was always there for continuous suggestions. Matteo Tiezzi from my same vision team, a talented guy. And all other people, that are so many - but all contribute to this great environment.

I would like to thank my friends and housemates at Casa Massari - for keeping me sane and balanced. Casa Massari has been my home and family away from home. I will look back very fondly at our time together.

Thanks to all my friends in my hometown Palermo, to make me feel always close. In particular to Dario Lo Castro, my twin at the time of university and still now.

Thanks to Tamara Cuasapaz, one of the greatest artists and person I have ever met, for constantly inspiring me by showing new ways for looking at the things of the world. She also designed Eymol's logo.

Thanks to my family for their continued support.

Thanks to Danilo Pileri, simply as a brother to me.

Thanks to my sister Francesca, who always gives me my most sincere laughs, and who together with Ciro gave me a wonderful niece, Elisa.

Thanks to my parents, Emanuele and Assunta, for providing me enough examples and for their simple advice to "do a good job".

Abstract

Visual attention is a crucial process for humans and foveated animals in general. The ability to select *relevant* locations in the visual field greatly simplifies the problem of *vision*. It allows a parsimonious management of the computational resources while catching and tracking coherences within the observed temporal phenomenon. Understanding the mechanisms of attention can reveal a lot about human intelligence. At the same time, it seems increasingly important for building intelligent artificial agents that aim at approaching human performance in real-world visual tasks. For this reasons, in the past three decades, many studies have been conducted to create computational models of human attention. However, these have been often carried over as the mere prediction of the *saliency map*, i.e. topographic map that represents conspicuousness of scene locations. Although of great importance and usefulness in many applications, this type of study does not provide an exhaustive description of the attention mechanism, since it misses to describe its temporal component.

In this thesis, we propose three models of scanpaths, i.e. trajectories of free visual exploration. These models share a fundamental idea: the evolution of the mechanisms of visual attention has been guided by fundamental functional principles. Scanpath models emerge as laws of nature, in the framework of mechanics. The approaches are mainly data-driven (bottom-up), defined on video streams and visual properties completely determine the forces that guide movements.

The first proposal (EYMOL) is a theory of free visual exploration based on the general Principle of Least Action. In the framework of analytic mechanics, a scanpath emerges in accordance with three basic functional principles: boundedness of the retina, curiosity for visual details and invariance of the brightness along the trajectories. This principles are given a mathematical formulation to define a potential energy. The resulting (differential) laws of motion are very effective in predicting saliency. Due to the very local nature of this laws (computation at each time step involve only a single pixel and its close surround), this approach is suitable for real-time application.

The second proposal (CF-EYMOL) expands the first model with the information coming from the internal state of a pre-trained deep fully convolutional neural network. A visualization technique is presented to effectively extract convolutional features (CF) activations. This information is then used to modify the potential field in order to favour exploration of those pixels that are more likely to belong to an object. This produces incremental results on saliency prediction. At the same time, it suggests how to introduce preferences in the visual exploration process through an external (top-down) signal.

The third proposal (G-EYMOL) can be seen as a generalisation of the previous works. It is completely developed in the framework of gravitational (G) physics. No special rule is described to define the direction of exploration, ex-

cept that the features themselves act as masses attracting the focus of attention. Features are given we assume they come from external calculation. In principle, they can also derive from a convolutional neural network, as in the previous proposal, or they can simply be raw brightness values. In our experiments, we use only two basic features: the spatial gradient of brightness and the optical flow. The choice, slightly inspired by the basic raw information in the earliest stage V1 of the human vision, is particularly effective in the experiments of scanpath prediction. The model also includes a dynamic process of inhibition of return defined within the same framework and which is crucial to provide the plus of energy for the exploration process. The laws of motion that are derived are integral-differential, as they also include sums over the entire retina. Despite this, the system is still widely suitable for real-time applications since only one step computation is needed to calculate the next gaze position.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"The Donkey Ride", Eva Gonzales, 1880. Impressionist artists aim at recreating the sensation in the eye that views the subject, rather than delineating the details.

The computational modeling of visual attention is at the crossroad of many and very different disciplines like computational neuroscience, physiology, information theory, psychology, machine learning [37]. This is due to the fact that attention is one of the most characteristic processes of human intelligence and deeply interwound with behavioural as well as physiological processes. This chapter of introduction aims to shortly introduce the terminology used throughout the text concerning the human visual system or the branch of computational modeling of visual attention and saliency. Most of the concepts or works mentioned here will be taken up along the text in more detail, when it is useful.

1.1 Human vision system

In this section we describe the main features of the human visual system. In our work of modelling visual attention processes, we have never been faithful to human biology; rather, we have defined general functional principles and researched how much these are related to the processes that take place in humans. At the same time, however, we have often taken inspiration from humans. Moreover, the data we used to validate the models performance refer to human behavioural processes. This makes it necessary to be familiar with some terms that describe the components of the human visual system and its different processes.

The human eye: structure and functions

The human eye is a sense organ that allows vision [51]. It has an almost spherical shape and it is composed of fluid enclosed into three layers (see Fig. 1.1). The outer layer is a fibrous tissue called sclera. At the front, however, this layer is transformed into a transparent disk that allows light rays to penetrate the eye: the cornea. The middle layer includes iris, the ciliar body and the choroid. Iris is the coloured part of the eye, just below the cornea. It contains muscles that control the size of the pupil, i.e. the opening at the centre of the iris. The ciliar body surrounds the lens. It includes muscles that control the refractive power of the lens and a vascular component that irrigates the front of the eye. In the choroid there is a rich layer of capillaries that irrigates the photoreceptors of the inner layer, the retina. The inner layer contains neurons sensitive to light and capable of transmitting visual signals. The space between the lens and the surface of the retina is filled with a thick, gelatinous liquid called vitreous humor, which makes up about 80% of the volume of the eye. In addition to maintaining the shape of the eye, this liquid contains cells that are intended to remove blood or other particles that may interfere with a clear vision. The retina has a circular area of about 1.5 millimeters in diameter in which the concentration of the cones is maximum. This area is called fovea and is placed more

medianally with respect to the optic nerve, also known as cranial nerve, which is the structure that transmits visual information from the retina to the brain.

Neural scheme

The visual system is the part of the central nervous system which gives organisms the ability to process visual detail. The part of the cerebral cortex that processes visual information is called *visual cortex*. Visual information coming from the eyes, reaches the visual cortex trough the geniculate nucleus in the thalamus. The part of the visual cortex that receives the sensory information from the thalamus is the primary visual cortex (V1). At the stage a velocity tag is associated to every major object of the observed scene. This is useful to predict object movements. V1 also performs edge detection to understand spatial organization and color changes.

The extrastriate areas V1, V2, V3, V4, V5 and V6 process information at a successive stage. V2 forwards and receives signals with V1, directly or trough the pulvinar (a group of nuclei located in the thalamus). Pulvinar is responsible for saccade and visual attention. V2 handles with illusory contours, determines depth by comparing left and right input and identifies the foreground distinguishment. V3 is involved in the computing of global motion of objects. V4 is the area dedicated to recognition of simple shapes. In V5 is where integration of local object motion with global motion takes place. integrates local object motion into global motion on a complex level. V6 works in conjunction with V5 for further motion analysis.

Both hemispheres of the brain contain a visual cortex and each of the two sides receives signals from the other.

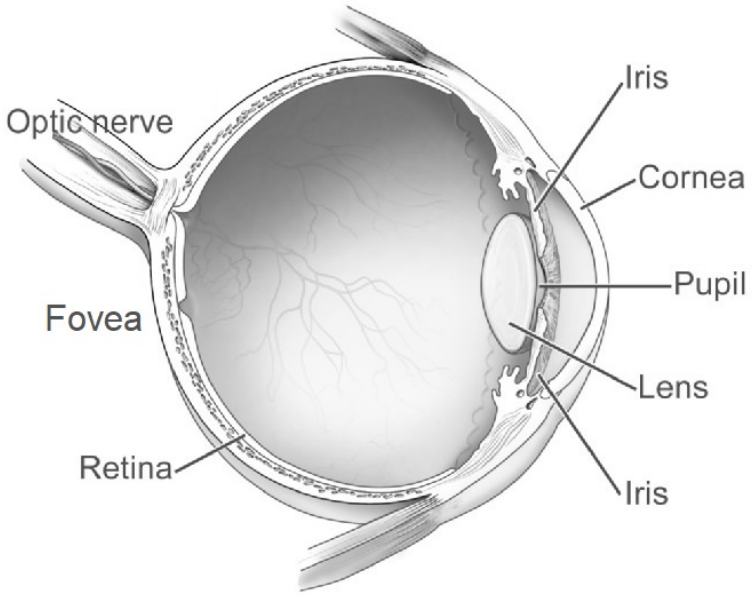


Figure 1.1: **Structure of the human eye.** This image shows the structure of an eye and the main components are labeled. The small dimple in the middle of the retina, close to the optic nerve, is the fovea. It is the center of the eye's sharpest vision and the location of most color perception.

Fixations and eye movements

Eye movements are an essential mechanism of human vision which allows to carry the fovea to each part of the image to be fixated upon and processed with high resolution. They play a fundamental role in stabilizing the gaze and are important in the task of recognition of patterns [65]. There are basically four types of eye movements [47, 48]: saccade, smooth pursuit, vergence movements and vestibulo-ocular movements. The functions of each of them are distinct and will be described as follows.

Saccades are rapid movements that modify the fixation point. They can be voluntary movements or can be performed in a reflective way. Saccades are also performed during a sleep phase called REM (Rapid Eye Movements) which is accompanied by dreams and physiological changes in heart rate, breathing and blood pressure. After the saccade has brought the gaze to a target, it takes about 200 milliseconds for the next movement to begin. This time interval identifies the so-called *fixations*, i.e. the maintaining of the visual gaze on a single location, during which information is acquired and the next shift is calculated with respect to the position of the fovea and the next target. This shift, or motor error, is converted into a motor control that will activate the extra-ocular muscles that will rotate the eye in the right direction. Their amplitude can vary a lot, depending on the task performed by a subject. For example, we need short saccades while reading and wide saccades while performing a task of free-visual exploration of a scene [62].

Smooth pursuit movements are much slower than saccades. They have the task of maintaining a moving stimulus aligned with the fovea. They are also voluntary movements in the sense that you can choose whether to track a target or not, but surprisingly only a few subjects are normally able to carry out a smooth pursuit in the absence of an actual visual moving target.

To align the fovea of each eye with targets positioned at different distances, humans perform *vergence movements*. They are reflexive movements (not voluntary) and, unlike the previous ones, in this case the eyes do not move in the same direction.

The *vestibulo-ocular movements* have the function of stabilizing the eyes with respect to the external environment to compensate for the movements of the head. Also vestibulo-ocular movements are reflexive.

Overt vs covert

Attention may be differentiated into *overt* versus *covert* orienting [76]. Overt attention can be observed in the form of eye movement to directly point eyes in the direction of a point in the visual field. Movements can be voluntary or reflexive. The overt mechanism of bringing the foveal area to bear on peripheral object to discern

their features is, in turn, supplemented by a covert - hidden - attention system that plays an important role in guiding eye movements. Studies show that observers can attend location in the periphery even while the eyes remain fixed [57].

Bottom-up vs top-down

Eye movements are an essential part of human vision as they drive the fovea and, consequently, selective visual attention toward a region of interest in space. Free visual exploration is an inherently stochastic process depending on image statistics (bottom-up) but also individual variability of cognitive and attentive state (top-down) [16, 56]. Thus, predictions about gaze trajectory and next fixation may be affected by variability and errors. Both low level and high level visual processing are involved in this task and interact during visual search.

However, there are scientific evidence that human visual attention is *data-driven*, at least partially. For example, in [1] authors show that saliency maximization is a strong behavioural drive that would prevail even during non-visual tasks. Exploratory behaviours were spatially similar to those of an explicit visual exploration task but they were, nevertheless, attenuated, reflecting slower visual sampling in this task. Exploration patterns are somehow independent on the task experimented, but they can occur with different velocity rate. In [61] it is shown as subjects tend to repeat sequences of scanpath when looking at the same picture (or pictures with slight modifications) at different times. Another clue is provided by newborns: despite their lack of knowledge of the world, they exhibit mechanisms of attention to extract relevant information from what they see [32]. Moreover, more precise studies show evidences that the very first fixations are highly correlated among adult subjects who are presented with a new input [70]. Human share a common *mechanism* that drive early fixations, while scanpaths diverge later under top-down influences.

1.2 Computational modeling of visual attention: overview

Feature integration theory

Many attempts have been made in the direction of modeling visual attention. Based on the feature integration theory of attention [72], Koch and Ullman in [44] assume that human attention operates in the early representation, which is basically a set of feature maps. They assume that these maps are then combined in a central representation, namely the *saliency map*, which drives the attention mechanisms.

More specifically, this model include different steps. First, several image pyramids are computed to enable computation of feature at different scales. Simple feature - intensity, color, and orientation - are extracted and the center-surround is calculated to quantify the contrast into the feature maps. This operation is to compare the average value of a center region to the average value of a surrounding region. The feature maps are summed up to create the so called *conspicuity maps*. Finally, the conspicuity maps are normalized, weighted and linearly combined to form the saliency map.

Saliency based models

The first complete implementation of the just described scheme was proposed by Itti *et al.* in [38] (see Fig. 1.2). In that paper, feature maps for color, intensity and orientation are extracted at different scales. Then center-surround differences and normalization are computed for each pixel. Finally, all this information is combined linearly in a centralized saliency map.

Several other models have been proposed by the computer vision community, in particular to address the problem of refining saliency maps estimation. They usually differ in the definition of saliency, while they postulate a centralized control of the attention mechanism through the saliency map. For instance, it has been claimed that the attention is driven according to a principle of information maximization [2, 11], by an opportune selection of surprising regions [36] or guided by a task [27]. A detailed description of the state of the art is given in [7].

Machine learning approaches have been used to learn models of saliency. Judd *et al.* [42] collected 1003 images observed by 15 subjects and trained an SVM classifier with low-, middle-, and high-level features. More recently, automatic feature extraction methods with convolutional neural networks achieved top level performance on saliency estimation [49, 74].

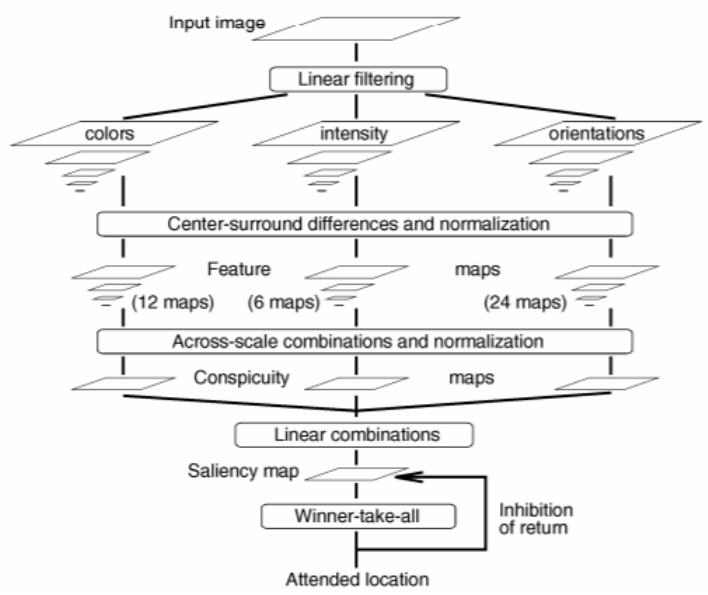


Figure 1.2: **Itti's model scheme.** This image is taken from the original paper [38]. It describes the scheme of computation from the input stimulus to the saliency map.

Baselines and human biases

Humans fixate the center of the images. The majority of human fixations appear to be next to the center and this is due to a viewing strategy by which subjects first inspect the image center, probably to rapidly gather a global view of the scene [52, 70] or because of the photographer bias to put interesting object in the middle of the scene [7, 8]. In the dataset MIT1003 [42], the average human saliency map from all 1003 images shows that 40% of fixations lie within the center 11% of the image; 70% of fixations lie within the center 25% of the image (see Fig. 1.3). A simple baseline Center, made using a Gaussian blob centered in the middle of the image, produces excellent results in saliency prodiction for many public datasets.

Other baselines have been proposed and demonstrate to be very effective. The *permutation control* [45] is calculated randomly sampling, for each image, fixations from a randomly-sampled image as our saliency map. This method allows us to capture observer and center biases that are independent of the image. The baseline *one human* is informative to understand how well a fixation map of one observer predicts the fixations of other observers. This can be calculated for each of the subject on a dataset and the final score is the average. Different subject can be more or less predictable so that this metric is associated with ranges for the prediction scores. The *random* baseline for saliency prediction can be calculated by assigning a random float number to each pixel of the visual field. This is the worse possible predictor.

Scanpath models

Most of the referred papers for saliency prediction share the idea that saliency is the product of a global computation. Some authors also provide scanpaths of image exploration, but to simulate them over the image, they all use the procedure defined by [44]. The *winner-take-all* algorithm is used to select the most salient location for the first fixation. Then three rules are introduced to select the next location: *inhibition-of-return*, *similarity preference*, and *proximity preference*. An attempt of introducing biological biases has been made by [53] to achieve more realistic saccades and improve performance.

A similar approach is defined in [61]. Authors use different image processing algorithms to extract maps of information content from generic images; local maxima are clustered to define regions of interest and scanpath order is obtain by ordering from the maximum to the minimum of those values.

Models of scanpaths that do not rely on a prior global computation of the saliency map have been proposed. Unfortunately, these works are often descriptive [54] or task specific [63] and the authors evaluate their models using static measures that do not consider the temporal evolution of the focus of the attention (e.g., they just

consider global improvements in the saliency map estimation or the average saccade length).

Moreover, all the approaches listed in this section have not been defined in the case of video streams and are difficult to extend to that case.

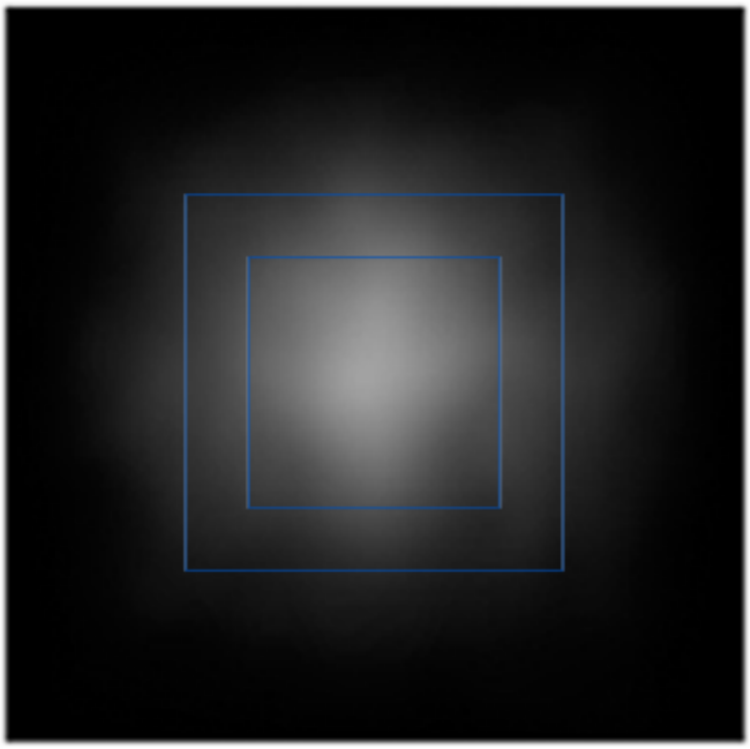


Figure 1.3: **Average of all fixations in MIT1003.** This image is taken from the original paper [42].

Chapter 2

Eymol

"Tree Drawings", Tim Knowles, 2005-2006. Trajectories are completely determined by external raw energies. The work was created by tying a brush to the branch of a tree.

In this chapter, we describe EYMOL (Eye movement laws), a model of visual attention that takes place in the earliest stage of vision, which we assume here to be completely data driven. We propose a theory of free visual exploration entirely formulated within the framework of physics and based on the general Principle of Least Action. The potential energy captures relevant features of the input space as well as crucial temporal properties, while the kinetic energy corresponds with the classic interpretation in analytic mechanics. Within this framework, bottom-up differential laws describing eye movements emerge in accordance with three functional principles. First, eye movements are bounded inside the definite area of the retina. Second, locations with high values of the brightness gradient are attractive. Third, trajectories are required to preserve the property of brightness invariance, which brings to fixation and tracking behaviours. To stress the model, we used a wide collection of images including basic features (pattern, sketch, fractals), noisy and low resolution, natural landscape, abstract (cartoon and line drawing), high level semantic content (social, affective, indoor), and more. In addition, the model was tested on a small video clip dataset. The derived differential equations are numerically integrated to simulate attentive scanpaths on both still images and videos. Results on saliency prediction benchmarks CAT2000, MIT300 and SFU, are presented to support the theory. Although the scores are higher when processing low semantic content, the model proves to be effective even in complex scenes. While alternatives exist for more accurate saliency map estimation, our approach provides a real-time model of visual search. It opens up the possibility of behavioural studies as well as of being integrated into learning processes of human-inspired machines. Despite the computation of the saliency maps only arises as a byproduct, the model outperforms all other classic [1] models in the literature [6, 9] in the task of saliency prediction.

2.1 Definitions

The *brightness* signal $b(t, x)$ can be thought of as a real-valued function

$$b : \mathbb{R}^+ \times \mathbb{R}^2 \to \mathbb{R} \tag{2.1}$$

where t is the time and $x = (x_1, x_2)$ denotes the position. The *scanpath* $x(t)$ over the visual input is defined as

$$x : \mathbb{R}^+ \to \mathbb{R}^2. \tag{2.2}$$

The scanpath will be also referred to as *trajectory* or *observation*.

[1] We indicate as *classic* those models that do not implement machine learning techniques to learn saliency directly from data.

2.2 Principles of visual attention

Three fundamental principles drive the model of attention. They lead to the introduction of the correspondent terms of the Lagrangian of the action.

i) *Boundedness of the trajectory*
Trajectory $x(t)$ is bounded into a defined area (retina). This is modeled by a harmonic oscillator at the borders of the image which constraints the motion within the retina[2]:

$$V(x) = k \sum_{i=1,2} \left((l_i - x_i)^2 \cdot [x_i > l_i] + (x_i)^2 \cdot [x_i < 0] \right) \tag{2.3}$$

where k is the elastic constant, l_i is the i-th dimension of the rectangle which represents the retina[3].

ii) *Curiosity driven principle*
Visual attention is attracted by regions with many details, that is where the magnitude of the gradient of the brightness is high. In addition to this local field, the role of peripheral information is included by processing a blurred version $p(t,x)$ of the brightness $b(t,x)$. The modulation of these two terms is given by

$$C(t,x) = b_x^2 \cos^2(\omega t) + p_x^2 \sin^2(\omega t), \tag{2.4}$$

where b_x and p_x denote the gradient w.r.t. x. Notice that the alternation of the local and peripheral fields has a fundamental role in avoiding trapping into regions with too many details.

iii) *Brightness invariance*
Trajectories that exhibit *brightness invariance* are motivated by the need to perform fixation. Formally, we impose the constraint

$$\dot{b} = b_t + b_x \dot{x} = 0.$$

This is in fact the classic constraint that is widely used in computer vision for the estimation of the optical flow [35]. Its soft-satisfaction can be expressed by the associated term

$$B(t,x,\dot{x}) = \left(b_t + b_x \dot{x}\right)^2. \tag{2.5}$$

Notice that, in the case of static images, $b_t = 0$, and the term is fully satisfied for trajectory $x(t)$ whose velocity $\dot{x}$ is perpendicular to the gradient, *i.e.* when the focus is on the borders of the objects. This kind of behaviour favours

[2]Here, we use Iverson's notation, according to which if p is a proposition then $[p] = 1$ if p=true and $[p] = 0$ otherwise

[3]A straightforward extension can be given for circular retina.

coherent fixation of objects. Interestingly, in case of static images, the model can conveniently be simplified by using the upper bound of the brightness as follows:

$$B(t,x,\dot{x}) = \dot{b}^2(t,x) = (\partial b_t + b_x\dot{x})^2 \leq 2b_t^2 + 2b_x^2\dot{x}^2 := \bar{B}(t,x,\dot{x}) \tag{2.6}$$

This inequality comes from the parallelogram law of Hilbert spaces. As it will be seen the rest of the paper, this approximation significantly simplifies the motion equations.

2.3 Least Action Principle

Visual attention scanpaths are modeled as the motion of a particle of mass m within a potential field. This makes it possible to construct the generalized action

$$S = \int_0^T L(t,x,\dot{x})\,dt \tag{2.7}$$

where $L = K - U$. K is the kinetic energy

$$K(\dot{x}) = \frac{1}{2}m\dot{x}^2 \tag{2.8}$$

and U is a generalized potential energy defined as

$$U(t,x,\dot{x}) = V(x) - \eta C(t,x) + \lambda B(t,x,\dot{x}). \tag{2.9}$$

Here, we assume that $\eta, \lambda > 0$. Notice that, while V and B get the usual sign of potentials, C comes with the flipped sign. This is due to the fact that, whenever it is large, it generates an attractive field. In addition, we notice that the brightness invariance term is not a truly potential, since it depends on both the position and the velocity. However, its generalized interpretation as a potential comes from considering that it generates a force field. Results on stationarity of the solution still hold in this case [29]. In order to discover the trajectory we look for a stationary point of the action in Eq. (2.7), which corresponds to the Euler-Lagrange equations

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}_i} = \frac{\partial L}{\partial x_i}, \tag{2.10}$$

where $i = 1,2$ for the two motion coordinates. More details about the principle of Least Action and its use in mechanics are given in the Appendix A. The right-hand term in (2.10) can be written as

$$\frac{\partial L}{\partial x} = \eta C_x - V_x - \lambda B_x. \tag{2.11}$$

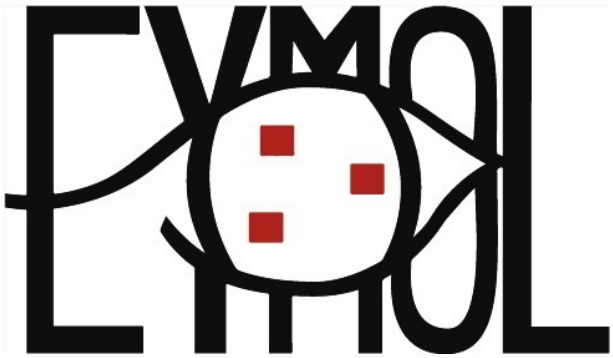
Likewise we have

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = m\ddot{x} - \lambda\frac{d}{dt}B_{\dot{x}} \tag{2.12}$$

so as the general motion equation turns out to be

$$m\ddot{x} - \lambda\frac{d}{dt}B_{\dot{x}} + V_x - \eta C_x + \lambda B_x = 0. \tag{2.13}$$

These are the general equations of visual attention. In what follows we give the technical details of the derivations. Throughout this work, the model defined by the equation 2.13 is referred to as the EYe MOvement Laws (EYMOL).



2.4 Variational laws of visual attention for dynamic scenes

In this section we compute in detail the differential laws of visual attention that describe the visual attention scanpath, as the Euler-Lagrange equations of the action functional 2.7.

First, we write down the partial derivatives of the different contributions w.r.t. x, in order to compute the exact contributions of 2.11. For the retina boundaries,

$$V_x = k\sum_{i=1,2}\big(-2\,(l_i - x_i)\cdot[x_i > l_i] + 2x_i\cdot[x_i < 0]\,\big). \tag{2.14}$$

The curiosity term in equation 2.4,

$$C_x = 2cos^2(\omega t)b_x\cdot b_{xx} + 2sin^2(\omega t)p_x\cdot p_{xx} \tag{2.15}$$

For the term of brightness invariance,

$$\begin{aligned} B_x &= \frac{\partial}{\partial x}\left(b_t + b_x\dot{x}\right)^2 \\ &= 2\left(b_t + b_x\dot{x}\right)\left(b_{tx} + b_{xx}\dot{x}\right). \end{aligned}$$

The authors thank Tamara Cuasapaz for the realization of the EYMOL logo.

Since we assume $b \in \mathcal{C}^2(t,x)$, the space of functions with 2 continuous derivatives, by the Schwarz's theorem[4], we have that $b_{tx} = b_{xt}$, so that

$$\begin{aligned} B_x &= 2\left(b_t + b_x\dot{x}\right)\left(b_{xt} + b_{xx}\dot{x}\right) \\ &= 2(\dot{b})(\dot{b}_x). \end{aligned} \tag{2.16}$$

We proceed by computing the contribution in (2.12). Derivative w.r.t. $\dot{x}$ of the brightness invariance term is

$$\begin{aligned} B_{\dot{x}} &= \frac{\partial}{\partial \dot{x}}\left(b_t + b_x\dot{x}\right)^2 \\ &= 2\left(b_t + b_x\dot{x}\right)b_x \\ &= 2(\dot{b})(b_x). \end{aligned} \tag{2.17}$$

So that, total derivative w.r.t. t can be write as

$$\frac{d}{dt}B_{\dot{x}} = 2\Big(\ddot{b}b_x + \dot{b}\dot{b}_x\Big) \tag{2.18}$$

We observe that $\ddot{b} \equiv \ddot{b}(t,x,\dot{x},\ddot{x})$ is the only term which depends on second derivatives of x. Since we are interested in expressing Euler-Lagrange equations in an explicit form for the variable $\ddot{x}$, we explore more closely its contribution

$$\begin{aligned} \ddot{b}(t,x,\dot{x},\ddot{x}) &= \frac{d}{dt}\dot{b} \\ &= \frac{d}{dt}(b_t + b_x\dot{x}) \\ &= \dot{b}_t + \dot{b}_x\cdot\dot{x} + b_x\cdot\ddot{x}. \end{aligned} \tag{2.19}$$

Substituting (2.19) in (2.18) we have

$$\begin{aligned} \frac{d}{dt}B_{\dot{x}} &= 2\Big((\dot{b}_t + \dot{b}_x\cdot\dot{x} + b_x\cdot\ddot{x})b_x + \dot{b}\dot{b}_x\Big) \\ &= 2\Big((\dot{b}_t + \dot{b}_x\cdot\dot{x})b_x + \dot{b}\dot{b}_x\Big) + 2(b_x\cdot\ddot{x})b_x. \end{aligned} \tag{2.20}$$

Given that, from (2.12) we get

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = m\ddot{x} - 2\lambda\Big((\dot{b}_t + \dot{b}_x\cdot\dot{x})b_x + \dot{b}\dot{b}_x + (b_x\cdot\ddot{x})b_x\Big). \tag{2.21}$$

Combining (2.11) and (2.21), we get the Euler-Lagrange equation

$$m\ddot{x} - 2\lambda\Big((\dot{b}_t + \dot{b}_x\cdot\dot{x})(b_x) + (\dot{b})(\dot{b}_x) + (b_x\cdot\ddot{x})b_x\Big) = \eta C_x - V_x - \lambda B_x. \tag{2.22}$$

[4]Schwarz's theorem states that, if $f : \mathbb{R}^n \to \mathbb{R}$ has continuous second partial derivatives at any given point in $\mathbb{R}^n$, then $\forall i,j \in \{1,...,n\}$ it holds $f_{x_ix_j} = f_{x_jx_i}$

In order to obtain explicit form for the variable $\ddot{x}$, we re-write equation 2.22 moving to the left all contributes which involve $\ddot{x}$, and to the right all other terms

$$\begin{aligned} m\ddot{x} - 2\lambda(b_x \cdot \ddot{x})b_x =& \eta C_x - V_x - \lambda B_x + 2\lambda((\dot{b}_t + \dot{b}_x \cdot \dot{x})(b_x) + (\dot{b})(\dot{b}_x)) \\ =& \underbrace{\eta C_x - V_x + 2\lambda(\dot{b}_t + \dot{b}_x \cdot \dot{x})(b_x)}_{A=(A_1,A_2)}. \end{aligned} \tag{2.23}$$

In matrix form, the equation 2.23 can be written as

$$\begin{pmatrix} m\ddot{x}_1 \\ m\ddot{x}_2 \end{pmatrix} - \begin{pmatrix} 2\lambda(b_{x_1}\ddot{x}_1 + b_{x_2}\ddot{x}_2)b_{x_1} \\ 2\lambda(b_{x_1}\ddot{x}_1 + b_{x_2}\ddot{x}_2)b_{x_2} \end{pmatrix} = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \tag{2.24}$$

which provides us with the system of two differential equations

$$\begin{cases} m\ddot{x}_1 - 2\lambda(b_{x_1}\ddot{x}_1 + b_{x_2}\ddot{x}_2)b_{x_1} = A_1 \\ m\ddot{x}_2 - 2\lambda(b_{x_1}\ddot{x}_1 + b_{x_2}\ddot{x}_2)b_{x_2} = A_2 \end{cases} \tag{2.25}$$

Grouping by same variable,

$$\begin{cases} (m - 2\lambda b_{x_1}^2)\ddot{x}_1 - 2\lambda(b_{x_1}b_{x_2})\ddot{x}_2 & = A_1 \\ -2\lambda(b_{x_1}b_{x_2})\ddot{x}_1 + (m - 2\lambda b_{x_2}^2)\ddot{x}_2 & = A_2 \end{cases} \tag{2.26}$$

We define

$$D = \begin{vmatrix} (m - 2\lambda b_{x_1}^2) & -2\lambda(b_{x_1}b_{x_2}) \\ -2\lambda(b_{x_1}b_{x_2}) & (m - 2\lambda b_{x_2}^2) \end{vmatrix} \tag{2.27}$$

$$D_1 = \begin{vmatrix} A_1 & -2\lambda(b_{x_1}b_{x_2}) \\ A_2 & (m - 2\lambda b_{x_2}^2) \end{vmatrix} \tag{2.28}$$

$$D_2 = \begin{vmatrix} (m - 2\lambda b_{x_1}^2) & A_1 \\ -2\lambda(b_{x_1}b_{x_2}) & A_2 \end{vmatrix} \tag{2.29}$$

By the Cramer's method we get differential equation of visual attention for the two spatial component, *i.e.*

$$\begin{cases} \ddot{x}_1 = \dfrac{D_1}{D} \\ \\ \ddot{x}_2 = \dfrac{D_2}{D} \end{cases} \tag{2.30}$$

Notice that, this raise to a further condition over the parameter λ. In particular, in the case values of $b(t, x)$ are normalized in the range $[0, 1]$, it imposes to chose

$$D \neq 0 \implies \lambda < \frac{m}{4} \tag{2.31}$$

In fact,

$$\begin{aligned} D &= (m - 2\lambda b_{x_1}^2)(m - 2\lambda b_{x_2}^2) - 4\lambda^2 (b_{x_1} b_{x_2})^2 \\ &= m\Big(m - 2\lambda(b_{x_1}^2 + b_{x_1}^2)\Big). \end{aligned} \tag{2.32}$$

For values of $b_x = 0$, we have that

$$D = m^2 > 0 \tag{2.33}$$

so that $\forall t$, we must impose

$$D > 0. \tag{2.34}$$

If $\lambda > 0$, then

$$\begin{aligned} m - 2\lambda(b_{x_1}^2 + b_{x_1}^2) &> 0 \\ \lambda &< \frac{m}{2(b_{x_1}^2 + b_{x_1}^2)}. \end{aligned} \tag{2.35}$$

The quantity on the right reaches its minimum at $\frac{m}{4}$, so that the condition

$$0 < \lambda < \frac{m}{4} \tag{2.36}$$

guarantees the well-posedness of the problem.

2.5 Energy balance analysis

In this section we discuss the dynamical behaviour of equation 2.13 by proposing the energy balance [5]. As usual, we introduce the adjoint variable

$$p = L_{\dot{x}},$$

as well as the Hamiltonian

$$H(t, x, p) = p \cdot \dot{x} - L(t, x, \dot{x}), \tag{2.37}$$

The introduction of p and H allows us to rewrite the Euler-Lagrange equations in the classic first-order canonical form

$$\begin{aligned} \dot{x} &= \frac{\partial H}{\partial p} \\ \dot{p} &= -\frac{\partial H}{\partial x}. \end{aligned} \tag{2.38}$$

[5] The authors thank Mattia Bongini for insightful discussion and suggestions for this analysis.

The derivative of the Hamiltonian turns out to be

$$\begin{aligned}\frac{dH}{dt} &= \frac{\partial H}{\partial t} + \frac{\partial H}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial H}{\partial p}\frac{\partial p}{\partial t} \\ &= \frac{\partial H}{\partial t} + \frac{\partial H}{\partial x}\frac{\partial H}{\partial p} - \frac{\partial H}{\partial p}\frac{\partial H}{\partial x} = \frac{\partial H}{\partial t}.\end{aligned} \tag{2.39}$$

This is a classic invariant that is expressed by the Poisson brackets $[H, H] = H_t$. Moreover, we have

$$\begin{aligned}\frac{\partial H}{\partial t} &= \frac{\partial}{\partial t}\left(\dot{x} \cdot p - L(t, x, \dot{x})\right) \\ &= \frac{\partial \dot{x}}{\partial t} \cdot p - \frac{\partial L}{\partial t} - p \cdot \frac{\partial \dot{x}}{\partial t} = -\frac{\partial L}{\partial t},\end{aligned} \tag{2.40}$$

which yields

$$\frac{dH}{dt} = -\frac{\partial L}{\partial t}. \tag{2.41}$$

We now investigate more closely the contribution given by the last term $\frac{\partial L}{\partial t}$. From (2.8) and (2.9), we compute

$$\frac{\partial L}{\partial t} = \frac{\partial C(t, x)}{\partial t} - \frac{\partial B(t, x, \dot{x})}{\partial t} \tag{2.42}$$

Notice that, whenever the input is a static image, we have $B_t(t, x, \dot{x}) = 0$ and, finally, $H_t = C_t$.

We can specifically identify the terms which continuously charge the system with new energy generated by the video. The stability of the process is favoured by the introduction of dissipation term according to

$$\bar{L}(t, x, \dot{x}) = e^{\theta t} L(t, x, \dot{x}). \tag{2.43}$$

with $\theta > 0$.

2.6 Parameters estimation with simulated annealing

Different choices of parameters lead to different behaviours of the system. In particular, weights can emphasize the contribution of curiosity or brightness invariance terms. To better control the system we use two different parameters for the curiosity term (2.4), namely η_b and η_p, to weight b and p contributions respectively. The best values for the three parameters $(\eta_b, \eta_p, \lambda)$ are estimated using the algorithm of simulated annealing (SA). This method allows to perform iterative improvements, starting from a known state i. At each step, the SA considers some neighbouring state j of the current state, and probabilistically moves to the new state j or stays on

the current state i. For our specific problem, we limit our search to a parallelepiped-domain D of possible values, due to theoretical bounds and numerical[6] issues. Distance between states i and j is proportional with a temperature T, which is initialized to 1 and decreases over time as

$$T_k = \alpha * T_{k-1}$$

, where k identifies the iteration step, and

$$0 << \alpha < 1.$$

The procedure is illustrated in detail in the Algorithm 1. The iteration step is repeated until the system reaches a state that is good enough for the application, which in our case is to maximize the NSS similarity between human saliency maps and simulated saliency maps.
Only a batch of a 100 images from CAT2000-TRAIN is used to perform the *SA* algorithm[7]. This batch is created by randomly selecting 5 images from each of the 20 categories of the dataset. To start the *SA*, parameters are initialized in the middle point of the 3-dimensional parameters domain D. The process is repeated 5 times, on different sub-samples, to select 5 parameters configurations. Finally, those configurations together with the average configuration are tested on the whole dataset, to select the best one.

Algorithm 1 In the psedo-code, P() is the acceptance probability and score() is computed as the average of NSS scores on the sample batch of 100 images.

```
1:  procedure SimulatedAnnealing
2:      Select an initial state i ∈ D
3:      T ← 1
4:      do
5:          Generate random state j, neighbor of i
6:          if P(score(i), score(j)) ≥ Random(0, 1) then
7:              i ← j
8:          end if
9:          T ← α * T
10:     while T ≥ 0.01
11: end procedure
```

The learning of the three parameters of the model by simulated annealing, being able to rely on batches of data on which to calculate the scores, is crucial. The

[6]Too high values for η_b or η_p produce numerically unstable and unrealistic trajectories for the focus of attention.

[7]Each step of the *SA* algorithm needs evaluation over all the selected images. Considering the whole dataset would be very expensive in terms of time.

behaviour of the model changes considerably with this choice, along with the performance on the saliency prediction task. It is worth mentioning, however, that along this text we will refer with the term "supervised" only to those deep learning models that learn a saliency model directly from human fixation data. The other models, including our proposals, will be referred as "unsupervised" or classic, according with the common terminology [6, 7].

2.7 Experiments

To quantitative evaluate how well our model predicts human fixations, we defined an experimental setup for saliency prediction both in images and in video. We used images from MIT1003 [42], MIT300 [41] and CAT2000 [8], and videos from SFU [31] eye-tracking dataset. Many of the design choices were common to both experiments; when they differ, it is explicitly specified.

Setup and data pre-processing

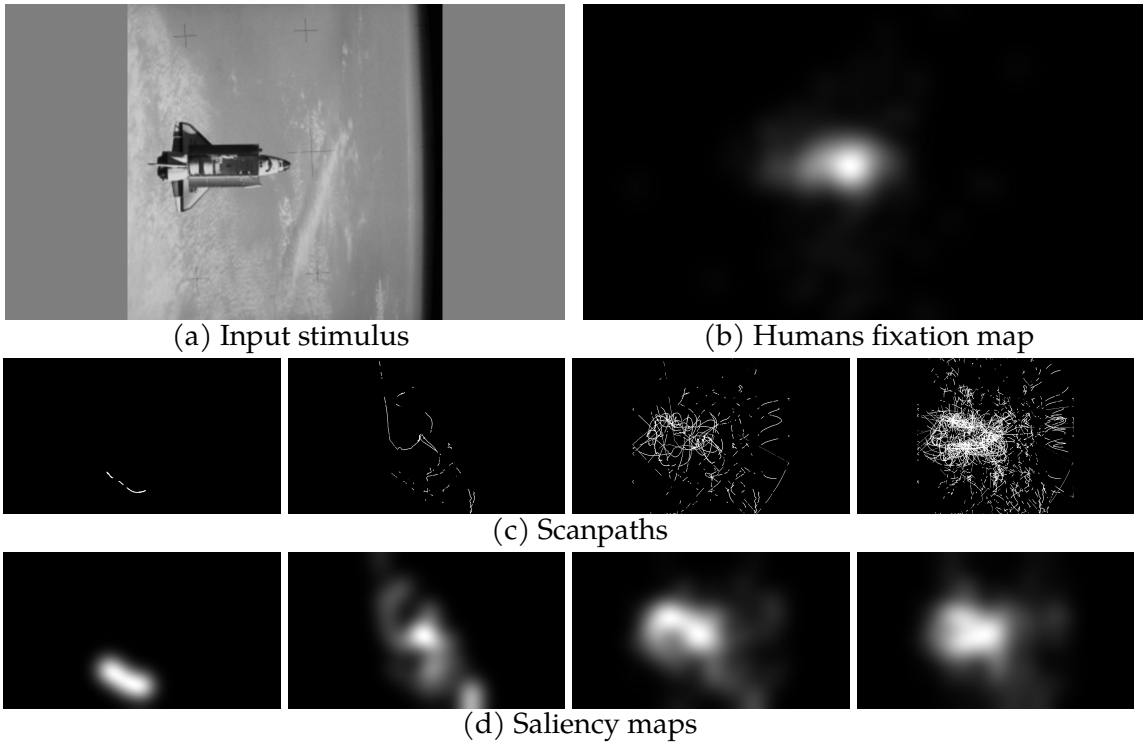
All input images are converted to gray-scale. Peripheral input p is implemented as a blurred versions of the brightness b. This blurred version is obtained by convolving the original gray-scale image with a Gaussian kernel. For the images only, an algorithm identifies the rectangular zone of the input image in which the totality of information is contained in order to compute l_i in equation 2.3. This is particularly useful in the case of images from the dataset CAT2000 since authors added gray bands to fill images to the fixed resolution 1920×1080 pixels. Finally both b and p are multiplied by a Gaussian blob centered in the middle of the frame in order to make brightness gradients smaller as we move toward periphery and produce a center bias.

Differently by many of the most popular methodologies in the state-of-the-art [11, 28, 38, 42, 74, 82], the saliency map is not itself the central component of our model but it can be naturally calculated from the visual attention laws in 2.13. The output of the model is a trajectory determined by a system of two second ordered differential equations, provided with a set of initial conditions. Since numerical integration of 2.13 does not raise big numerical difficulties, we used standard functions of the python scientific library *SciPy* [40].

Saliency map is then calculated by summing up the most visited locations during a sufficiently large number of virtual observations (see Fig. 2.1). For images, we collected data by running the model 199 times, each run was randomly initialized almost at the center of the image and with a small random velocity, and integrated for a running time corresponding to 1 second of visual exploration. For videos, we

collected data by running the model 100 times, each run was initialized almost at the center of the first frame of the clip and with a small random velocity.

Model that have some blur and center bias on the saliency map can improve their score with respect to some metrics. A grid search over *blur radius* and *center* parameter σ have been used, in order to maximize AUC-Judd and NSS score on the training data of CAT2000 in the case of images, and on SFU in case of videos.



(a) Input stimulus (b) Humans fixation map

(c) Scanpaths

(d) Saliency maps

Figure 2.1: **How to create a saliency map with EYMOL.** We simulate a task of free-viewing. In 2.1c is shown the output of the EYMOL model corresponding to 1, 10, 50 and 199 virtual observers exploration over image 2.1a. Optimized saliency maps in 2.1d are obtained by convolving images in 2.1c with Gaussian kernel.

Dataset description

- **MIT1003** [42]. This dataset contains 1003 natural indoor and outdoor scenes. They are sampled with variable sizes, where each dimension is in 405-1024px. The database contains 779 landscape images and 228 portrait images. Fixations of 15 human subjects are provided for 3 seconds of free-viewing observation.

- **MIT300** [41]. This dataset contains 300 natural indoor and outdoor scenes. They are sampled with variable sizes, where each dimension is in 405-1024px. Test data is kept private and scores are provided by the MIT Saliency Team [12]. Human fixations are collected with the same experimental conditions used for the collection of MIT1003, and for this reason MIT1003 can be used as a training set in predicting saliency for MIT300.

- **CAT 2000** [8]. We select the publicly available portion of this dataset, that contains 2000 images from 20 different categories. Stimuli include basic features (basic patterns, sketches, fractals), noisy and low resolution images, natural landscapes, abstract pictures (cartoon and line drawing), high level semantic contents (social, affective, indoor), and more. The resolution of the images is 1920x1080 px. Saliency maps are provided for each image. Maps are calculated with data on 18 different human subject free-viewing exploration.

Saliency metrics

Different metrics are used in order to evaluate a saliency predictor. There is an open discussion in the scientific community about what is the best way to evaluate models. Pros and cons about different metrics have been investigated in [12]. The most common saliency metrics are:

- **Normalized Scanpath Saliency (NSS).** It is measured as the mean value of the normalized saliency map at fixation locations. NSS ≤ 0 indicates that the model performs no better than picking a random position on the map.

- **Area Under the ROC Curve (AUC).** The saliency map is treated as a binary classifier of fixations. The true positive (tp) rate (proportion of saliency map values above threshold at fixation locations) and the false positive (fp) rate (proportion of saliency map values above threshold at non-fixated pixels) are calculated at different threshold values to create the ROC curve. Random saliency maps have a score of AUC $= 0.5$.

- **Kullback-Leibler (KL) divergence.** The KL divergence is measure of distance between the distributions of saliency values at human versus random eye positions. Suppose we are given information about N huamn fixations. For a

given model, saliency is estimate at a human fixation point $x_{i,human}$ and at a random point $x_{i,random}$. Saliency magnitude is normalized in the range $[0, 1]$. The histograms in q bins of this values are calculated and we indicate with H_k and R_k the franction of point in the k-th bin for human and random points respectively. Finally the KL-divergence between the two histograms is calculated as

$$\mathrm{KL} = \frac{1}{2} \sum_{k=1}^{q} \left(H_k \log \frac{H_k}{R_k} + R_k \log \frac{R_k}{H_k} \right). \tag{2.44}$$

- **Similarity Measure (SIM).** It is also called histogram intersection and measures the similarity between two different saliency maps when viewed as distributions. When the distributions are identical, then SIM $= 1$.

- **Earth Mover's Distance (EMD).** It is a measure of distance between two probability distributions. It expresses how much transformation one distribution would need to undergo to match another. Informally, distributions are interpreted as two different ways of piling up a certain amount of dirt over the region D and the EMD is the minimum cost of turning one pile into the other. Identical distributions have EMD $= 0$.

- **Linear correlation coefficient (CC).** This metric is used to compare the relationship between two images in applications like image registration, object recognition or disparity measurement. CC $= 0$ indicate that the two maps are uncorrelated. It is also known as Pearson's linear coefficient. Given a saliency map S and a human fixations map F (a map with 1 at fixations point and 0 elsewhere), the linear correlation coefficient is defined as

$$CC(S, F) = \frac{\sum_{x,y} \left(F(x,y) - \mu_F\right) \left(S(x,y) - \mu_S\right)}{\sqrt{\sigma_G^2 \sigma_S^2}} \tag{2.45}$$

  at each fixation location (x, y).

AUC and NSS are considered the most robust metrics and often the results are reported only for these two. However, it is advisable to provide the results for all these metrics as they can provide useful and different information. Qualitatively, better models are those whose provide good scores with respect to as many of the metrics as possible.

Results

Two versions of the the model have been evaluated. The first version V1 implementing brightness invariance in the approximated form (2.6), the second version V2 implementing the brightness invariance in its exact form derived in section 2.4. Models V1 and V2 have been compared on the MIT1003 and CAT2000-TRAIN datasets,

since they provide public data about fixations. Parameters estimation have been conducted independently for the two models and the best configuration for each one is used in this comparison. Results are statistically equivalent (see Tab. 2.1 and 2.2) and this proves that, in the case of static images, the approximation is very good and does not cause loss in the score. For further experiments we decided to use the approximated form V1 due to its simpler form of the equation that also reduces time of computation.

Model V1 has been evaluated in two different dataset of eye-tracking data: MIT300 and CAT2000-TEST. In this case, scores were officially provided by MIT Saliency Benchmark Team [12]. Further considerations about the used metrics are provided in [13]. Table 2.3 and 2.4 shows the scores of our model compared with five other popular method [11, 28, 38, 42, 74], which have been selected to be representative of different approaches. Despite its simplicity, our model reaches best score in half of the cases and for different metrics.

We evaluated our model in a task of saliency prediction also on dynamic scenes with the dataset SFU [31]. The dataset contains 12 clips and fixations of 15 observers, each of them have watched twice every video. Table 2.5 provides a comparison with other four model. Also in this case, despite of its simplicity and even if it was not designed for the specific task, our model competes well with state-of-the-art models. Our model can be easily run in real-time to produce an attentive scanpath. In some favourable cases, it shows evidences of tracking moving objects on the scene.

Model version	CAT2000-TRAIN AUC	NSS
V1 (approx. br. inv.)	0.8393 (0.0001)	1.8208 (0.0015)
V2 (exact br. inv.)	0.8376 (0.0013)	1.8103 (0.0137)

Table 2.1: **EYMOL V1 vs V2 (CAT2000-TRAIN).** Comparison between EYMOL implemented with the approximated (V1) and the exact form (V2) for the brightness invariance term. Between brackets is indicated the standard error.

Model version	MIT1003 AUC	NSS
V1 (approx. br. inv.)	0.7996 (0.0002)	1.2784 (0.0003)
V2 (exact br. inv.)	0.7990 (0.0003)	1.2865 (0.0039)

Table 2.2: **EYMOL V1 vs V2 (MIT1003).** Comparison between EYMOL implemented with the approximated (V1) and the exact form (V2) for the brightness invariance term. Between brackets is indicated the standard error.

Model	MIT300 AUC	SIM	EMD	CC	NSS	KL
Itti-Koch [38], implem. by [33]	0.75	0.44	4.26	0.37	0.97	**1.03**
AIM [11]	0.77	0.40	4.73	0.31	0.79	1.18
Judd Model [42]	0.81	0.42	4.45	**0.47**	**1.18**	1.12
AWS [28]	0.74	0.43	4.62	0.37	1.01	1.07
eDN [74]	**0.82**	0.44	4.56	0.45	1.14	1.14
EYMOL	0.77	**0.46**	**3.64**	0.43	1.06	1.53

Table 2.3: **Results on saliency prediction (MIT300).** Results are provided by MIT Saliency Benchmark Team [12]. The models are sorted chronologically. In bold, the best results for each metric and benchmarks.

Model	CAT2000-TEST AUC	SIM	EMD	CC	NSS	KL
Itti-Koch [38], implem. by [33]	0.77	0.48	3.44	0.42	1.06	**0.92**
AIM [11]	0.76	0.44	3.69	0.36	0.89	1.13
Judd Model [42]	0.84	0.46	3.60	0.54	1.30	0.94
AWS [28]	0.76	0.49	3.36	0.42	1.09	0.94
eDN [74]	**0.85**	0.52	2.64	0.54	1.30	0.97
EYMOL	0.83	**0.61**	**1.91**	**0.72**	**1.78**	1.67

Table 2.4: **Results on saliency prediction (CAT2000).** Results are provided by MIT Saliency Benchmark Team [12]. The models are sorted chronologically. In bold, the best results for each metric and benchmarks.

	SFU Eye-Tracking Database EYMOL	Itti-Koch [38]	Surprise [36]	Judd Model [42]	HEVC [78]
Mean AUC	0.817	0.70	0.66	0.77	0.83
Mean NSS	1.015	0.28	0.48	1.06	1.41

Table 2.5: **Results on saliency prediction on videos (SFU).** Scores are calculated as the mean of AUC and NSS metrics of all frames of each clip, and then averaged for the 12 clips.

2.8 Discussion

In this chapter we investigated an attention mechanisms that emerges in the early stage of vision, which we assume completely data-driven, and compared it with human data. The proposed model consists of differential equations, which provide a real-time model of scanpath. These equations are derived in a generalized framework of Least Action, which nicely resembles related derivations of laws in physics. A remarkable novelty concerns the unified interpretation of curiosity-driven movements and the brightness invariance term for fixation and tracking, that are regarded as mechanisms that jointly contribute to optimize the acquisition of visual information. Experimental results on both images and videos datasets for saliency prediction are promising, especially if we consider that the proposed theory offers a truly model of eye movements, whereas the computation of the saliency maps only arises as a byproduct.

This approach seems to be very suitable with theories of feature extraction that are still expressed in terms of variational-based laws of learning [30, 55] that are based on the concept of temporal coherence. The brightness invariance term, in fact, pushes to find solutions that preserve visual information of brightness along the trajectory. The consistency of the video stream selected along the scanpath makes it a good candidate for the search of temporal coherences useful for understanding the scene and for learning. This aspect has not been investigated yet, so that we leave as suggestion for future works.

In chapters 4, we will investigate the presented model in the case of behavioural data, not only in terms of saliency maps, but also by comparing actual generated scanpaths with human data and to discover temporal correlations. We anticipate that, despite the very good results on saliency prediction, the model definition suffers of some drawbacks that penalize its results on scanpath prediction. For example, it is worth notice at this point that the derived laws are very local and this scarcely reflect the actual human vision structure. A model does not need to necessary emulate the human vision system in its organization, but especially in the case of peripheral vision a mechanism that aggregates information coming from locations which are far from the actual fixation point are very important. Approximation done at this stage (with a sinusoidal function to alternate the brightness b with its blurred version p) is an *ad hoc* solution and does not work enough effectively.

The just presented model has been recently included in a survay [6, 9] that analyzes the problem of saliency prediction in the era of deep learning. According to the author's analysis, EMYOL [80] and BMS [82] are the best among the classic models, *i.e.* those models that do not implement machine learning techniques to learn saliency directly from a ground truth of human fixations.

Chapter 3

CF-Eymol

"Architectonic", Ljubov' Sergeevna Popova, 1917. Cubist artists attempted to show objects as the mind, not the eye, perceives them.

In some cases, deep learning models automatically develop a *inherent model of attention*, even when they trained for a different task (for example, image classification). In this chapter, we show how to visualize attention maps from fully convolutional neural networks. We also quantitatively measure how much this form of attention developed by an artificial system is similar to attention in humans while freely exploring static scenes. Finally, we show how this can be integrated with the scanpath model presented in the previous chapter to bring incremental results on the task of saliency prediction along with the property of being attracted by the main objects of the scene.

3.1 Inherent visual attention in deep convolutional neural networks

Recently, in the strand of explainability of deep learning, efforts have been made to understand what deep models *actually* learn. In the case of CNNs, some methods [66, 83] allow to visualize internal activation and understand which locations of the original input were crucial for the system response. In particular, authors remove the fully-connected layer before the final output and replace it with global average pooling followed by a fully-connected softmax layer. Class-specific activation maps are then obtained by averaging feature maps from the last convolutional tensor with the weights of the correspondent class.

While commonly used as a regularization technique for training, a closer investigation of the role of the global pooling term [83] reveals that it actually allows the convolutional neural network to develop *localization ability*. This property is successfully used in a number of different task.

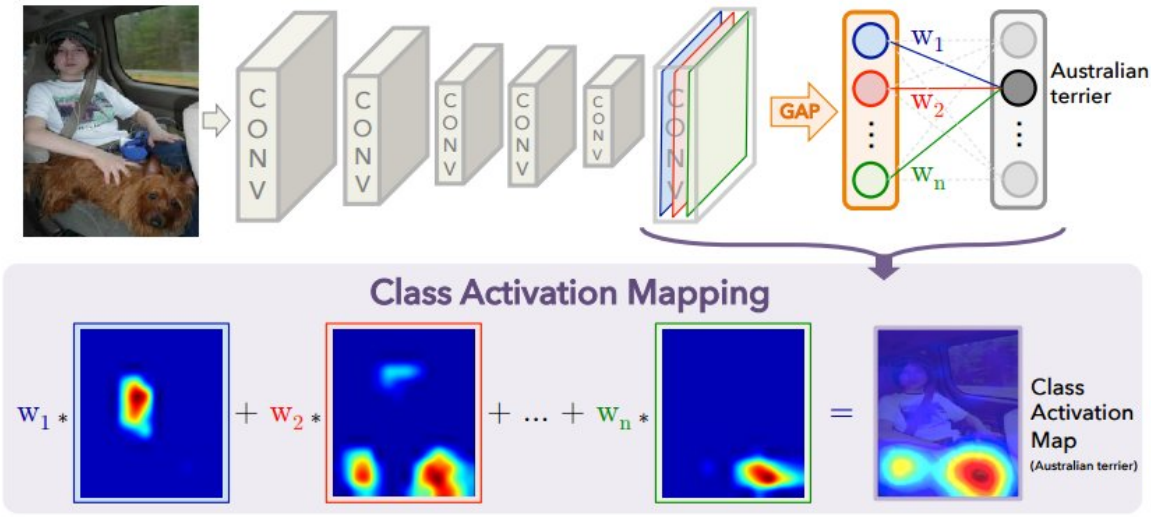


Figure 3.1: **Class activation maps (CAM).** This picture is taken from the original paper [83] presenting the idea. Authors revisit the global average pooling layer and show how it enables CNN to have remarkable class specific localization ability, despite being trained on image-level labels.

3.2 Visualization technique

Guided by the same principle in [83] but not being interested in one particular class, we claim that semantic maps obtained by averaging the activation of the units in the last convolutional layer are good predictors of the human fixations distribution. In this section, we make a quantitative analysis of how well this maps predict human fixations, even if they belong to a model that have been trained for a different task.

In our experiments, we used an instance of the model described in [69], pre-trained for classification on the ImageNet benchmark [1] (see Tab.3.1 for more architectural specifications). The architecture is described in detail in Fig. 3.1. Since it is a fully-convolutional model (until the last level before the final softmax) it is suitable for the purpose of constructing semantic maps [83].

For a given input image, let $f_k(x)$ represents the activation of unit k in the last convolutional layer "pool" at spatial location $x = (x_1, x_2)$, and $k = 1, ..., K \in \mathbb{R}$. Then, we can indicate the result of global average pooling as

$$F_k = \sum_x f_k(x). \tag{3.1}$$

For each class c of the dataset, the input of the softmax is

$$\sum_k w_k^c F_k, \tag{3.2}$$

where w_k^c is the weight corresponding to class c for unit k. Notice that w_k^c expresses the importance of F_k for the class c. In [83], class-specific activation maps are defined for each class as

$$M^c(x) = \sum_k w_k^c f_k(x). \tag{3.3}$$

For our purpose of building a model of free-viewing, we are not interested in any specific class. Then, we can average the activations by setting

$$w_k^c = K^{-1}, \forall k. \tag{3.4}$$

We remove the subscript c and indicate with

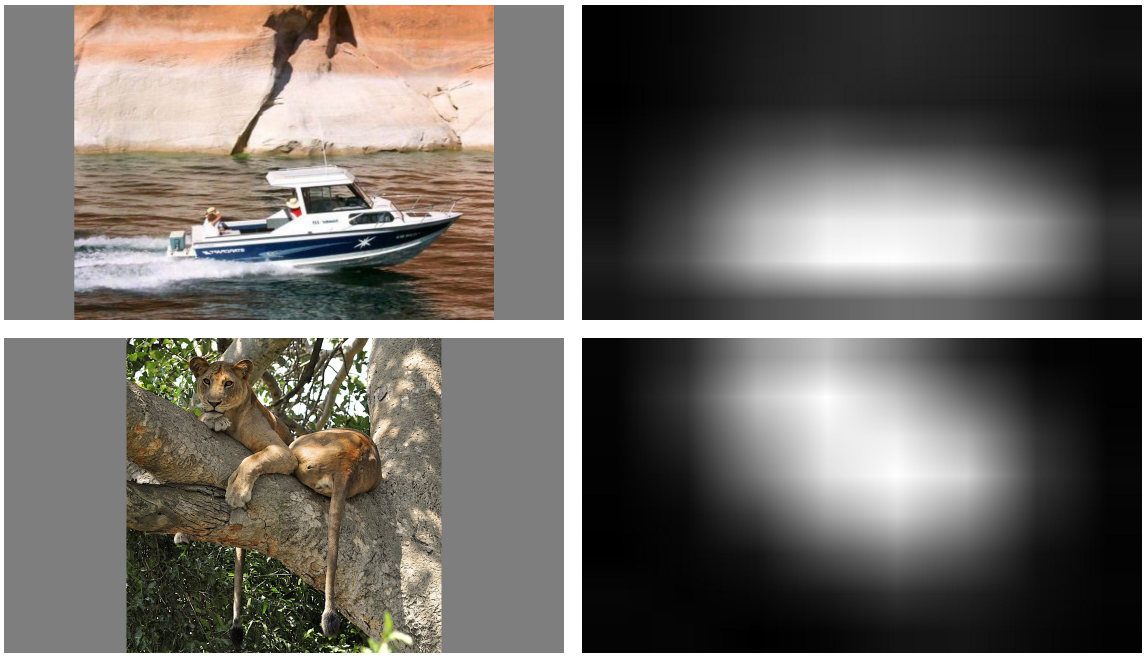
$$M(x) = \frac{1}{K} \sum_k f_k(x) \tag{3.5}$$

the map of the average activations of the features on the last convolutional layer, defined on each spatial location x. Examples of this maps are given in Fig. 3.2. Notice that the resulting maps have a much lower resolution than the original input, because of the many pooling operations. In our experiments, these maps have been resized with cubic interpolation to the original input size in order to obtain a map M defined on each pixel.

[1] http://image-net.org

Type	patch size / stride	input size
conv	$3 \times 3/2$	$299 \times 299 \times 3$
conv	$3 \times 3/1$	$149 \times 149 \times 32$
conv padded	$3 \times 3/1$	$147 \times 147 \times 32$
pool	$3 \times 3/2$	$149 \times 149 \times 64$
conv	$3 \times 3/1$	$73 \times 73 \times 64$
conv	$3 \times 3/2$	$71 \times 71 \times 80$
conv	$3 \times 3/1$	$35 \times 35 \times 192$
$3\times$ Inception	See fig.5 in [69]	$35 \times 35 \times 288$
$5\times$ Inception	See fig.6 in [69]	$17 \times 17 \times 768$
$2\times$ Inception	See fig.7 in [69]	$8 \times 8 \times 1280$
pool	8×8	$8 \times 8 \times 2048$
linear	logits	$1 \times 1 \times 2048$
softmax	classifier	$1 \times 1 \times 1000$

Table 3.1: **Inception-v3.** Architecture specifications of the model of CNN described in [69].



(a) Stimulus (b) M

Figure 3.2: **Convolutional feature (CF) activation map M.** In column 3.2a, examples of images from CAT2000 [8]. In column 3.2b the correspondent map M obtained from the pre-trained instance of inception-v3 [69].

Results on saliency prediction

The regions of the image with the highest activation are those more likely to contain the most relevant information for the task, i.e. the most salient object of the scene. This behaviour can be qualitatively observed in Fig. 3.2. If the hypothesis that humans direct their gaze towards objects is true [22], we may expect to find correlation between these maps and the humans fixations maps. We have quantitatively evaluated this hypothesis on CAT2000 [8] by measuring how well M predicts human fixations maps. Performance improve by optimizing maps with blurring and histogram matching [13]. Scores are reported in Tab. 3.2. From now on, we will refer to this model as the Convolutional Feature map (CF). The CF model performs better then classic models and, especially for the NSS metrics, competes with state-of-the-art deep learning models as well. This result is impressive if we consider that no training or refining techniques have been applied but maps are the output of a simple visualization technique.

Please notice that some very similar approaches already exist in the literature. In [50] the authors use fixation data to learn weight for the activation map and optimize saliency prediction. In this case, authors chose not to use a fully convolutional model and this probably penalize their performance.

Model version	Maps optimization	CAT2000 AUC	CAT2000 NSS
CF	-	0.80 (0.001)	1.177 (0.046)
CF	center bias	**0.844 (0.001)**	1.168 (0.009)
CF	center bias, hist. match.	0.834 (0.001)	**1.684 (0.085)**
Itti-Koch	[38], implem. by [33]	0.77	1.06
AIM	[11]	0.76	0.89
Judd Model	[42]	0.84	1.30
AWS	[28]	0.76	1.09
eDN	[74]	0.85	1.30
DeepFix	[49]	0.87	2.28
SAM	[17]	**0.88**	**2.38**

Table 3.2: **Results on saliency prediction (CAT2000).** Between brackets is indicated the standard error.

3.3 Attention guided by convolutional features

Integrating convolutional features activation CF with EYMOL

In the previous chapter, the dynamic model of visual attention EYMOL is derived by three functional principles. We repeat definitions here as well to let each chapter of the working being self-contained. The three basic principles are the following. First, eye movements are required to be bounded inside the definite area of the retina,

$$V(x) = k \sum_{i=1,2} \left((l_i - x_i)^2 \cdot [x_i > l_i] + (x_i)^2 \cdot [x_i < 0] \right). \tag{3.6}$$

Second, locations with high values of the brightness gradient are attractive. This gives the potential term

$$C(t,x) = b_x^2 \cos^2(\omega t) + p_x^2 \sin^2(\omega t). \tag{3.7}$$

Finally, trajectories are required to preserve the property of brightness invariance, which brings to fixation and tracking behaviors. This is guaranteed by the soft satisfaction of the constraint

$$B(t,x,\dot{x}) = \left(b_t + b_x \dot{x}\right)^2. \tag{3.8}$$

This makes it possible to construct the generalized action

$$S = \int_0^T L(t,x,\dot{x})\, dt \tag{3.9}$$

where $L = K - U$. K is the kinetic energy

$$K(\dot{x}) = \frac{1}{2} m \dot{x}^2 \tag{3.10}$$

and U is a generalized potential energy defined as

$$U(t,x,\dot{x}) = V(x) - \eta C(t,x) + \lambda B(t,x,\dot{x}). \tag{3.11}$$

By the Principle of Least Action, the true path of a mass m within the defined potential fields is given by the Euler-Lagrange equations

$$m\ddot{x} - \lambda \frac{d}{dt} B_{\dot{x}} + V_x - \eta C_x + \lambda B_x = 0. \tag{3.12}$$

The model of scanpath defined by equation 3.12 is referred to as EYMOL. This equations can be numerically integrated to simulate processes of free visual exploration on images and videos.

Nevertheless, this model in (3.12) is too naive. It fails in those categories that contain high level semantic content, for example those picture containing faces, writings, emotional content. The principles 3.6, 3.7 and 3.8 are, in fact, very local: they depend on the fixated pixel value and its small surround and fail in capturing properties at the object-level.

This can be solved by adding to the system a external (top-down) signal which suggests what of the regions are more likely to contain an object and, even more, the main object of the scene [14]. For this reason, we propose to extend the model in equation 3.12 by adding the information carried by the convolutional feature activation maps defined in equation 3.5. The versatility of the framework allows us to do so simply by modifying the potential energy 3.11 as follows

$$\bar{U}(t, x, \dot{x}) = U(t, x, \dot{x}) - \gamma M(x), \tag{3.13}$$

that brings to the new model

$$m\ddot{x} - \lambda \frac{d}{dt} B_{\dot{x}} + V_x - \eta C_x + \lambda B_x - \gamma M_x = 0, \tag{3.14}$$

where M_x corresponds to the spatial derivative of M. The signal M carries the information of how much the pixel $x = (x_1, x_2)$ belongs to a salient object. From now on, we will refer to the EYMOL model enriched with convolutional features CF with the acronym CF-EYMOL.

Notice that, this method for integrating an external signal is very general. The external signal is calculated independently. It may provide information other than the preference of object-like figures. For example, the mechanism may be asked to prefer eye movements that minimize the distance to a certain target (for a tracking task), or it may be asked to favour fixations on those locations that contain a certain feature of interest (for a search task).

3.4 Experiments

Dataset description

- **CAT 2000** [8]. We select the publicly available portion of this dataset, that contains 2000 images from 20 different categories. Stimuli include basic features (basic patterns, sketches, fractals), noisy and low resolution images, natural landscapes, abstract pictures (cartoon and line drawing), high level semantic contents (social, affective, indoor), and more. The resolution of the images is 1920x1080 px. Saliency maps are provided for each image. Maps are calculated with data on 18 different human subject free-viewing exploration.

Results on saliency prediction

As EYMOL, also for CF-EYMOL the saliency map is not itself the central component of the model. The output of the model is a trajectory determined by a system of two second ordered differential equations 3.14, provided with a set of initial conditions. Saliency map is calculated as byproduct by summing up the most visited locations after a certain number of virtual observations.

Table 3.3 reports scores for saliency prediction on the CAT2000 dataset. Saliency maps are obtained by running the model 199 times, each run was randomly initialized almost at the center of the image with a small random velocity, and integrated for a running time corresponding to 1 second of visual exploration. The addition of convolutional features CF brings improvements in performance.

Improvements are more evident in the case in which the number trials is dramatically reduced from 199 to 10, as it can be seen in Table 3.4. This can be explained by the fact that the peripheral prior provided by the convolutional features CF is more crucial when only a few fixations are granted to the system. It is worth notice that, this last configuration still runs in real-time in a average equipped personal computer and its performance is comparable or better than the *one-human* baseline.

Model version	Maps optimization	CAT2000 AUC	CAT2000 NSS
CF	center bias, hist. match.	0.834 (0.001)	1.684 (0.085)
EYMOL	blur	0.838 (0.001)	1.810 (0.014)
CF-EYMOL	blur	**0.843 (0.001)**	**1.822 (0.064)**
Itti-Koch	[38], implem. by [33]	0.77	1.06
AIM	[11]	0.76	0.89
Judd Model	[42]	0.84	1.30
AWS	[28]	0.76	1.09
eDN	[74]	0.85	1.30
DeepFix	[49]	0.87	2.28
SAM	[17]	**0.88**	**2.38**

Table 3.3: **Results on saliency prediction (CAT2000).** Saliency map summarize results for 199 virtual observations. Different virtual observations are obtained by small variations on the initial conditions of the differential system. Between brackets is indicated the standard error.

Model version	Maps optimization	CAT2000 AUC	CAT2000 NSS
EYMOL	blur	0.805 (0.001)	1.428 (0.031)
CF-EYMOL	blur	**0.821 (0.001)**	1.524 (0.040)
{One-human}	[12]	0.76	**1.54**

Table 3.4: **Results on saliency prediction (CAT2000).** Saliency map summarize results for 10 virtual observations. Different virtual observations are obtained by small variations on the initial conditions of the differential system. Models between curly brackets are baseline. Between brackets is indicated the standard error.

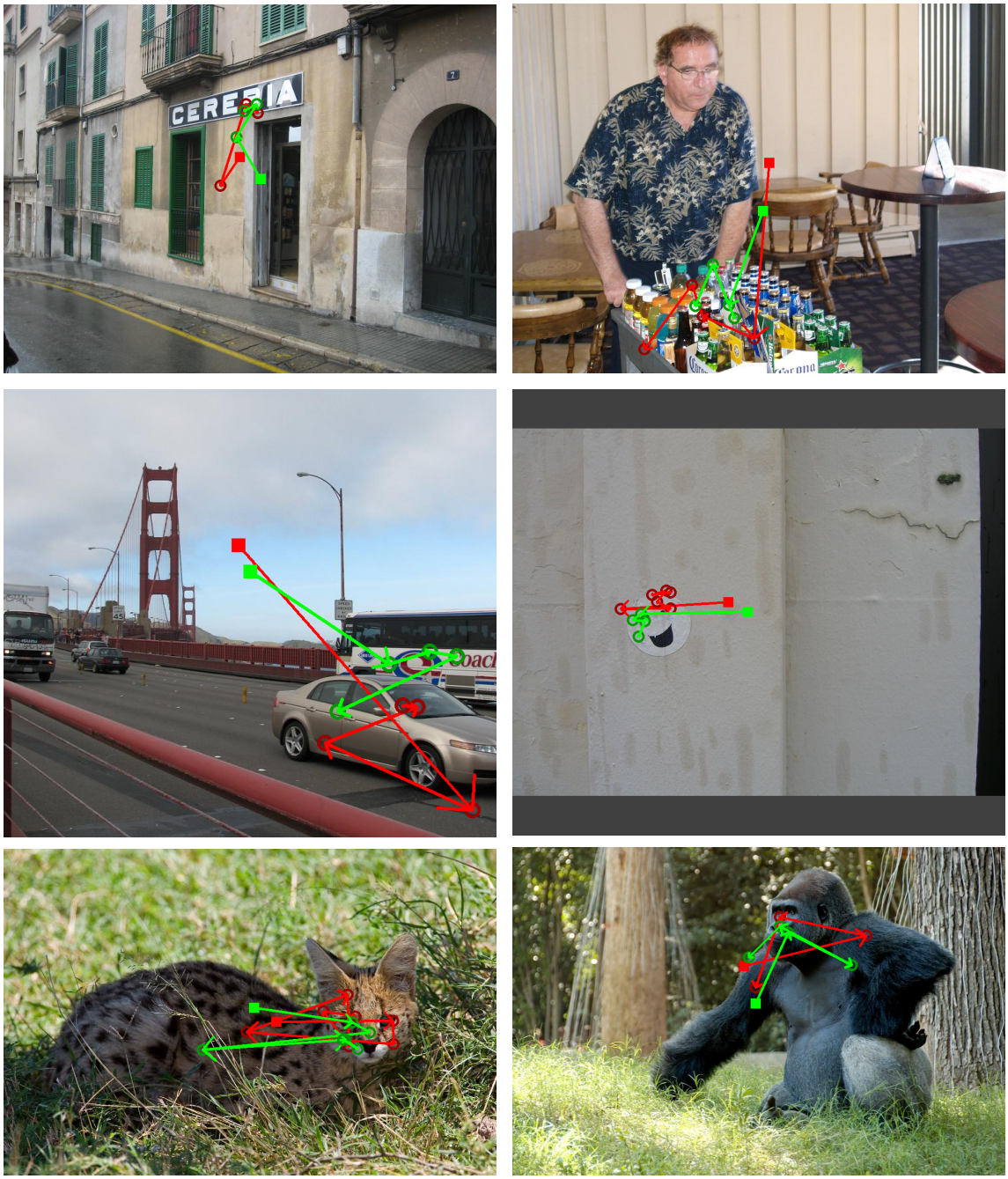


Figure 3.3: **Simulated scanpaths with CF-EYMOL.** This figures show a qualitative comparison of the scanpaths simulated with CF-EYMOL and human scanpaths. Simulated scanpaths are drawn in red, human scanpath in green. The starting point is marked with a square and the arraws represents saccades and their directions.

3.5 Connections with the Yarbus' theory

Depending on the task one person is engaged, the distribution of the points of fixation varies correspondingly, depending on the information needed to solve that specific task. This is clearly because different parts of information are localized in different parts of the image. In the seminal work of Yarbus [79], images containing complex objects or scenes are presented to a subject which is asked to answer a question. In one case, for example, Yarbus presents to the subject a picture of a group of people in a room. The distribution of fixation points is very different as the task varies: when the subject is asked to "estimate the material circumstances of the family in the portrait", his fixations fall very much above the clothes of the people, while he tends to look at their faces when he is asked to "give the ages of the people shown in the picture". In the same work, a big number of example is presented that produce similar results. Further and independent successive scientific investigations [5, 20, 59] have confirmed in more solid setup the Yarbus' hypothesis.

In this chapter, the potential field guiding the CF-Eymol trajectories of visual attention is enriched with the contribution of the activations in the last convolutional layer of a CNN. This produce an enriched potential field. It turns out, see Fig. 3.3, that trajectories are very different then the trajectories produced by its *naive* version Eymol. The main object of the scene tend to be visited in the very first fixations and the trajectories tend, then, to orbit around it. The reader may agree that this type of exploratory behaviour is functional to the task of object classification, as it is classically posed in machine learning. This task requires, in fact, that the algorithm assigns to each image a semantic label that describes the main object in the scene. Usually there are no ambiguous cases in the data, i.e. for a human it is very clear which of the labels is to be assigned to a certain image because it is related to the object that occupies the scene most or is positioned in the center of the frame.

While Eymol can be considered a completely data-driven model, i.e. based on a bottom-up information embedded in the distribution of the spatial frequences in the image, it is not the same story for CF-Eymol. The features developed by Inception-v3 convolutional neural network, which modify the potential field through their activations, depend on the task that CNN was required to solve. In other words, the basic exploratory mechanism remains the same but the enriched field changes the resulting saliency and order of priority in which different locations must be expected.

Going into the treatment of the phenomenon of visual attention in all its complexity goes beyond the objective of this thesis. It has been stated several times that the interest is to model the bottom-up component. At the same time, however, this chapter shows the versatility of the framework. The result of this chapter is a plausible instance in which the Yarbus' theory holds providing a practical modelling of the intent that in Yarbus' theory influences not only the positions of the fixations,

but also the order and their duration. In future works it is desirable to explore this hypothesis in a more solid way, verifying how, as the task varies, the distributions of fixations on the same scene vary as well.

3.6 Discussion

In this chapter we have shown that an inherent model of visual attention is present in deep convolutional neural networks that are trained for a different task. Our proposal is a very simple visualization techniques of the last convolutional layer of the considered deep learning model. The experimental results show that the convolutional features activations leads good human saliency predictors.

However, the main contribution in this chapter is to integrate the information brought by these maps with the bottom-up differential model of eye-movements EYMOL, defined in the previous chapter 2, with the final purpose of simulating visual attention scanpaths of CF-EYMOL, guided by convolutional feature activations. The proposed integration enriches the eye movement model thanks to the additional peripheral information that comes from the convolutional filters, as well as the information about salient objects. This integration determines an incremental performance in the task of saliency prediction.

Real-time performance on an average personal computer makes the approach suitable for real-time application, for example to improve systems of video surveillance [64, 71], where latency of the system is a very important factor, and convolutional features can be trained for special classes of interest. This direction has not been investigated and we leave it as suggestion for future works.

A final qualitative consideration is about the type of scanpaths that emerge with this new model CF-EYMOL. As show in some figures of example (see Fig. 3.3), they are very much oriented toward the main object of the scene. This is an encouraging results, since it means that external signal can be effectively used to influence the behaviour of the model. This is very reminiscent of what happens in humans, where fixations are highly dependent on the task that the subject has in mind at the time of data collection.

Chapter 4

G-Eymol

"States of Mind I: Those who Leave", Umberto Boccioni, 1911. The focal point of the picture is provided by movement itself.

In this chapter we propose G-EYMOL, which can be seen as a generalisation of the previous two works. The "G" stands for gravitational, since it is completely developed in the framework of gravitational physics. Differently by EYMOL and its modified version CF-EYMOL, here no specific principles are defined, except that the features themselves act as masses attracting the focus of attention. Features are defined outside the motion model and in principle, they can also derive from a convolutional neural network, like in CF-EYMOL or a similar approach. In our experiments, we use only two basic features: the spatial gradient of brightness and the optical flow. The choice, slightly inspired by the basic raw information in the earliest stage V1 of the human vision, is particularly effective in the experiments of scanpath prediction. Also inspired from biology, the model also includes a dynamic process of inhibition to return. It is defined within the same framework and it is prove to provide the plus of energy for making the exploration process not vanish. The laws of motion that are derived are integral-differential, as they also include sums over the entire retina. Despite this, the system is still widely suitable for real-time applications.

4.1 Salient features

We consider a video defined over the retinal domain $\mathcal{D} = \mathcal{R} \times \mathcal{T}$, where $\mathcal{R} \subset \mathbb{R}^2$ is the retina while $\mathcal{T} \subset \mathbb{R}$ is the temporal basis. The trajectory of the focus of attention is driven by a virtual mass $\mu : \mathcal{D} \to \mathbb{R}$ which yields a gravitational field associated with *relevant visual* features. This mass arises from the sum of different contributions. In this paper, we consider two different visual features as sources of virtual masses:

- Let $b : \mathcal{D} \to \mathbb{R}$ be the brightness of the video. It generates the *spatial gradient of the brightness*
$$\mu_1 = \alpha_1 \|\nabla_x b\|, \tag{4.1}$$
with $\alpha_1 \in \mathbb{R}^+$, so as the virtual mass $\mu_1(x,t)$ is available for all $(x,t) \in \mathcal{D}$. Clearly, $\mu_1(x,t)$ carries information about edges and, generally speaking, it reveals the presence of details in the video.

- Let $v : \mathcal{D} \to \mathbb{R}$ be the *optical flow*, that is the velocity field at any $(x,t) \in \mathcal{D}$. It generates the virtual mass
$$\mu_2 = \alpha_2 \|v\|, \tag{4.2}$$
with $\alpha_2 \in \mathbb{R}^+$, that characterizes moving areas in the retina.

In doing so, the focus of attention is either controlled by details, that are typically characterized by significant values of μ_1 or by moving objects, that produce significant values of μ_2. Basically, details and movements turn out to attract the correspon-

dent virtual masses, so as the process of focus of attention is translated into *gravitational attraction of attention* (see Fig. 4.1-A). More generally, the underlying idea of virtual masses can also be extended to the case in which attention is controlled by understanding processes. In this case, one can generate virtual masses by means of the visual features of a convolutional neural network.

4.2 Gravitational field

Now, let us consider a distribution of virtual mass μ. In case it degenerates to a single distributional mass concentrated in x, so as $\mu(y,t) = \delta(y-x)$, we can associate the trajectory of the focus of attention $a(t)$ with the potential

$$G(a-x) = -\frac{1}{2\pi}\log(\|a-x\|). \tag{4.3}$$

We can promptly see that $-G$ is the Green function of the Laplacian Δ, that is

$$\Delta G(a-x) = \delta(\|a-x\|). \tag{4.4}$$

Then, the gravitational field is simply $e = -\nabla G$, that is

$$e(a-x) = \frac{1}{2\pi}\frac{x-a}{\|x-a\|^2}. \tag{4.5}$$

Notice that, as one expects, the focus of attention a is attracted by the virtual mass at position x according to $\|e\| \propto 1/r$, where $r = \|x-a\|$. This is in fact the kind of radial dependency that a gravitational field is expected to exhibit in two-dimensional spaces. A straightforward way of understanding the reason for such a radial dependency is based on Gauss' theorem. If we consider a circle $\mathcal{C}$ with center x and radius $\|x-a\|$, then the flux of e on $\partial\mathcal{C}$ turns out to be

$$\int_{\partial\mathcal{C}} -\nabla G(a-x)\cdot\frac{a-x}{\|a-x\|}da = -\frac{1}{2\pi}\int_{\partial\mathcal{C}}\frac{1}{\|a-x\|}da = -1.$$

On the other hand, from Eq. (4.4), we have

$$-\int_{\mathcal{C}}\nabla\times\nabla G(y-x)dy = -\int_{\mathcal{C}}\delta(\|y-x\|) = -1,$$

and, finally, the consistency of the field e given by Eq. (4.5) arises from the divergence Gauss' theorem

$$\int_{\mathcal{C}}\nabla\times\nabla G(y-x)dy = \int_{\partial\mathcal{C}}\nabla G(a-x)\cdot\frac{a-x}{\|a-x\|}da.$$

This clearly explains the choice of Green function (4.3), along with the corresponding field, that are different with respect to 3D mass distributions. Notice that the

-1 in Eq. (4.5) is due to the attraction that unitary particle a receives from mass $\delta(y-x)$. Given any virtual mass μ, that comes from visual features as previously explained, we can construct the overall field by

$$E(a(t)) = -\frac{1}{2\pi}\int_{\mathcal{R}} dx \frac{a(t)-x}{\|a(t)-x\|^2}\mu(x,t). \tag{4.6}$$

Hence we can compactly can re-write $E(a)$ using the convolution operator as follows

$$E(a(t)) = -(e * \mu)(a(t)) \tag{4.7}$$

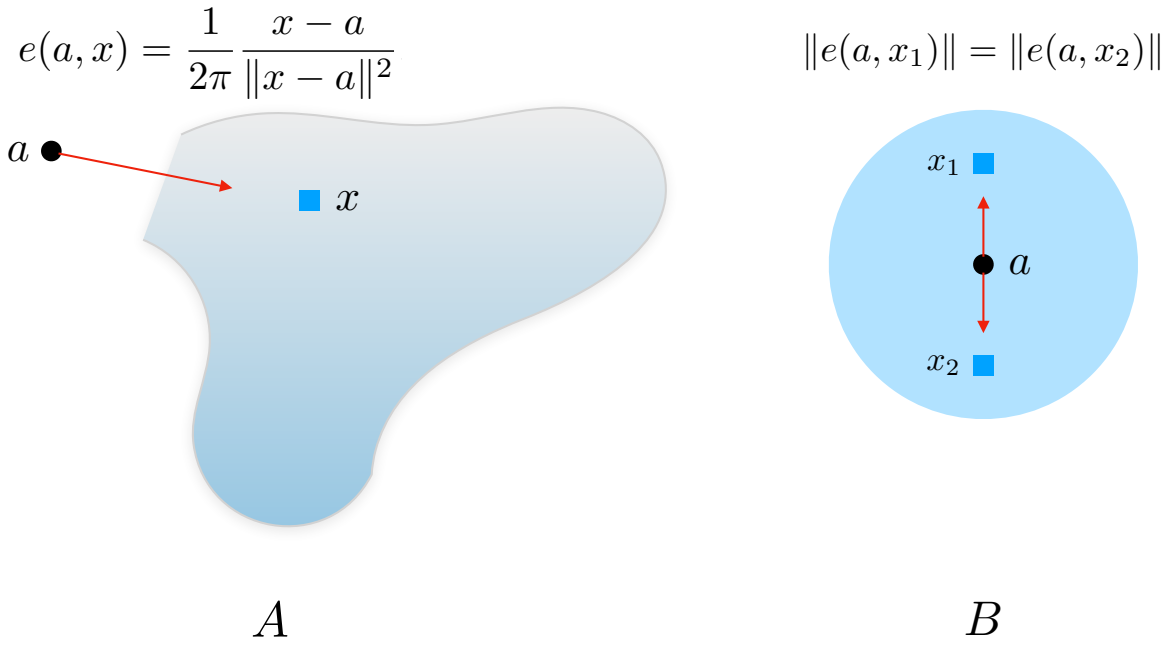


Figure 4.1: **Gravitational masses.** (A) The focus of attention can be regarded as an elementary mass which is attracted by the distributed mass in the drawn regions. (B) The gravitational effect of a symmetric mass on the focus of attention is null.

4.3 Inhibition of return

In humans, after a reflexive shift of attention towards the source of stimulation, there is an inhibition to remain in the same location. This inhibitory effect is referred to as *Inhibition-of-Return* (IOR) and it was early described in [60]. Interestingly, IOR is shown to have dedicated circuits in the human visual system. The advantage introduced by IOR is to encourage orienting attention towards unexplored locations and facilitate a complete exploration of the scene. We can promptly see that in order to provide an appropriate interpretation of IOR we need to enrich the given notion of virtual mass μ. In particular, when it comes from details in the retina, it might yield remarkable mass that constantly attracts the focus of attention. While regions that are dense of details are worth exploring, the idea behind IOR is exactly that of inhibiting those regions after awhile, so as to permit elsewhere exploration. This problem especially arises in the case of still images, where those regions can represent a trap for the focus of attention trajectory. Hence it is convenient to introduce the *inhibitory function* $I : \mathcal{D} \to [0,1]$ which is expected to return values $I(x,t) \simeq 0$ at the beginning of the visit of point x and $I(x,t) \simeq 1$ when the neighborhood of x has already been satisfactorily explored. We can model the inhibitory function I by

$$\frac{\partial I(x,t)}{\partial t} + \beta I(x,t) = \beta g(x - a(t)), \tag{4.8}$$

where

$$g(u) = e^{-\frac{u^2}{2\sigma^2}}$$

and

$$0 < \beta < 1.$$

The inhibitory function I can properly be used to transform virtual mass $\mu_1(x,t)$ into $\mu_1(x,t)(1 - I(x,t))$, whereas it is reasonable not to inhibit virtual masses μ_2 coming from motion to allow smooth tracking behaviour. Hence, we have

$$\mu(x,t) = \mu_1(x,t)(1 - I(x,t)) + \mu_2(x,t). \tag{4.9}$$

We are now ready to write the Newtonian differential equation of the focus of attention trajectory[1]. We have

$$\ddot{a}(t) + \lambda \dot{a}(t) + (e * \mu)(t, a(t)) = 0. \tag{4.10}$$

Here, there is also dumping term $\lambda \dot{a}(t)$ which prevents strong obscillations and makes the overall dynamics closer to human scanpath. Notice that the trajectory defined by $a(t)$ comes out from the coupling of equations (4.8), (4.9) and (4.10).

[1]Please notice that this is just the formulation on our problem setting of the second law of dynamics that states that the acceleration of an object produced by force is directly proportional to the magnitude of the force, in the same direction as the net force, and inversely proportional to the mass of the object. In formulas, $a = \frac{F}{m}$, where in our case $m = 1$

These equations are numerically integrated to simulate eye movements in all the experiments of this paper. Fixations and saccades emerge from the given gravitational laws (4.10). The overall dynamics also includes smooth pursuit, a phenomenon which is observed in humans: when tracking a moving target, humans do perform smooth movement on the gaze to be aligned with that target. It is appealing that all human eye movements emerge from the same differential formula.

4.4 Saliency and inhibitory function

Once the trajectory $a(t)$ is determined, $g(x - a(t))$ returns the degree of saliency of x at time t. Its averaging in a closed time interval $[0, T]$, with $T \in \mathbb{R}^+$, is

$$S_\theta(x) = \frac{\theta}{e^{\theta T} - 1} \lim_{T\to\infty} \int_0^T dt\, e^{\theta(T-t)} g(x - a(t)). \tag{4.11}$$

carried out by using the density $\frac{\theta}{e^{\theta T}-1} e^{\theta(T-t)}$ expresses the saliency in x. Clearly, as $\theta \to 0$ then $S_0(x)$ is the corresponding saliency with uniform temporal density and collapse to the same procedure described in Fig. 2.1. The growth of θ leads to values of the saliency which emphasize the recent (close to T) behaviour of the trajectory. Now, we will see that the saliency $S_\theta(x)$ is strictly related with the Laplace transform of the inhibitory function $I(x, t)$.

Given $I(x, t)$ we assume that it admits the Laplace transform

$$\hat{I}(x, s) = \int_0^\infty dt\, e^{-st} I(x, t).$$

The following theorem states formally the mentioned connection with the saliency $s_\theta(x)$.

Theorem 1. *For any point x of the retina $\mathcal{R}$ the saliency $S_\theta(x)$ and the Laplace transform of the inhibitory function $\hat{I}(x, \theta)$ are related by*

$$S_\theta(x) = \theta \frac{1 + \beta}{\beta} \hat{I}(x, \theta) \tag{4.12}$$

Proof. From Eq. (4.8), when taking the Laplace transform of both sides with argument $s\theta$ we get

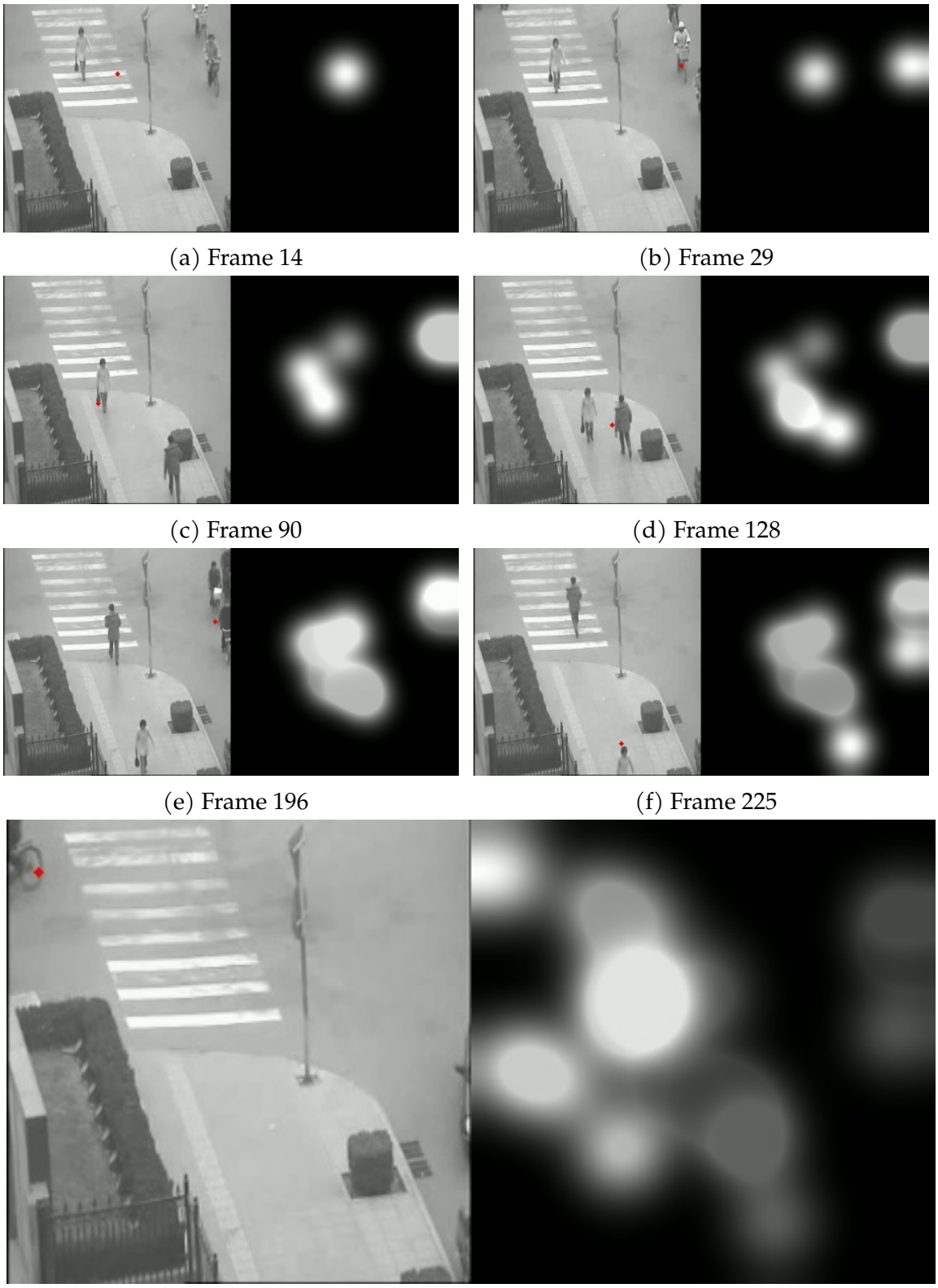
$$\begin{aligned} &s\hat{I}(x, \theta s) - I(x, 0) + \beta \hat{I}(x, \theta s) \\ &= \beta \lim_{T\to\infty} e^{-\theta s T} \int_0^T dt\, e^{\theta s(T-t)} g(x - a(t)). \end{aligned}$$

Since $I(x, 0) = 0$, for $s = 1$ we get

$$\begin{aligned} \hat{I}(x, \theta) + \beta \hat{I}(x, \theta) &= \beta \lim_{T\to\infty} e^{-\theta T} \int_0^T dt\, e^{\theta(T-t)} g(x - a(t)) \\ = \beta \lim_{T\to\infty} (e^{\theta T} - 1) e^{-\theta T} &\frac{\int_0^T dt\, e^{\theta(T-t)} g(x - a(t))}{e^{\theta T} - 1} = \frac{\beta}{\theta} S_\theta(x), \end{aligned}$$

from which we immediately draw the conclusion stated by Eq. (4.12). □

The condition $\hat{I}(x,\theta) \propto S_\theta(x)$ formally states that the inhibitory process is expressed by a function that is proportional to the saliency: areas with more saliency are those subject to highest inhibition of the virtual mass μ_1. This is straightforward if we consider that the salient areas correspond to the most visited ones and, as it is defined, the inhibition of return is applied on the same visited areas. Nevertheless, the analysis quantifies exactly the relationship between these two phenomena.



(a) Frame 14 (b) Frame 29

(c) Frame 90 (d) Frame 128

(e) Frame 196 (f) Frame 225

(g) Frame 374

Figure 4.2: **Example of inhibition in a video.** This figures show how the inhibition function evolves (right) while exploring a scene (left). The red dot indicates the actual point of focus of attention simulated with the proposed model. Please notice that the inhibition function decays over time and location which were highly inhibited, then became interesting again.

4.5 Energy balance analysis

Now we carry out an energy-based analysis of the dynamical process that provides a strong motivation on the adoption of described inhibitory process. For the i-th coordinate of $a(t)$, with $i \in \{1,2\}$, Eq. (4.10) can be rewritten multiplying it by $\dot{a}_i$, leading to

$$\dot{a}_i \ddot{a}_i + \lambda \dot{a}_i^2 + \frac{1}{2\pi} \int_{\mathcal{R}} dx \mu(x) \frac{a_i(t) - x_i}{\|a(t) - x\|^2} \dot{a}_i = 0.$$

If we integrate over $[0, t]$ we get

$$\int_0^t d\frac{1}{2}\dot{a}_i^2 + \lambda \int_0^t d\tau \dot{a}_i^2 + \frac{1}{2\pi} \int_0^t \int_{\mathcal{R}} dx d\tau \mu(x) \frac{a_i(t) - x_i}{\|a(t) - x\|^2} \dot{a}_i = 0.$$

Now we have

$$\frac{d}{dt} \int_{\mathcal{R}} dx \mu(x) \log\|a - x\| = \int_{\mathcal{R}} dx \mu(x) \frac{a_i(t) - x_i}{\|a(t) - x\|^2} \dot{a}_i + \int_{\mathcal{R}} dx \dot{\mu}(x) \log\|a - x\|,$$

and, therefore, we get

$$M(t) = K(t) - K(0) + U(t) - U(0) + \lambda D(t), \tag{4.13}$$

where

$$\begin{aligned} K(t) &:= \frac{1}{2}(\dot{a}_1^2 + \dot{a}_2^2) \\ U(t) &:= \frac{1}{2\pi} \int_{\mathcal{R}} dx \mu(x) \log\|a - x\| \\ D(t) &:= \int_0^t d\tau \dot{a}_i^2 \\ M(t) &:= \int_0^t \int_{\mathcal{R}} dx d\tau \dot{\mu}(x) \log\|a - x\| \end{aligned}$$

Here, K is the *kinetic energy* while U can be interpreted as a *potential energy* in case of constant mass. The term D represents the dissipated energy that forces the vanishing of the focus of attention trajectory. On the opposite, the term M injects energy thanks to the modification of the virtual mass. For reasons that will become clear from the statement of the following theorem, the term M is referred to as the *inhibitory energy*.

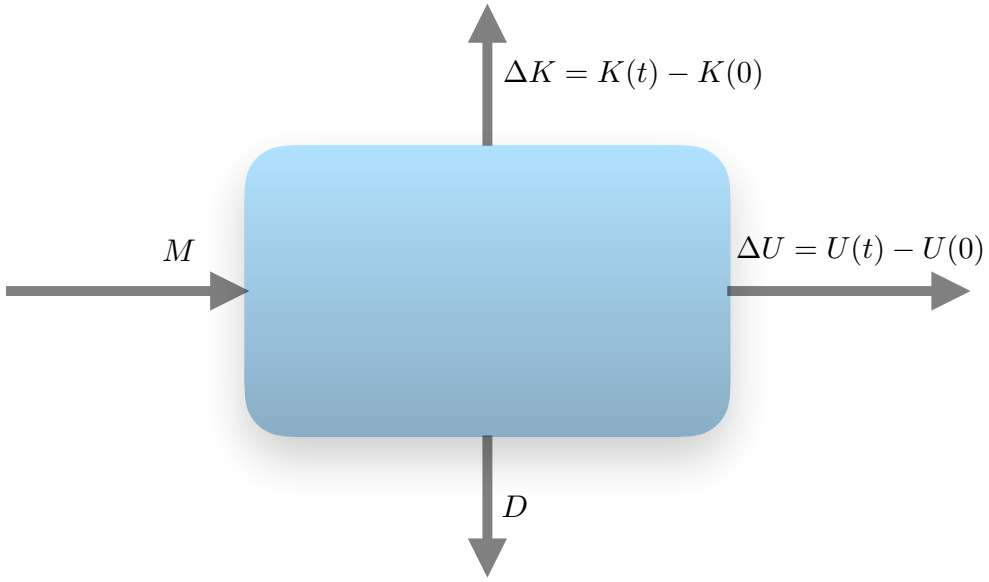


Figure 4.3: **Energy balance.** The energy variation $\Delta(U + K) = \Delta U + \Delta K$ along with the dissipated energy D is balanced by the injection of inhibitory energy M.

The overall balancing energy process is depicted in Fig. 4.3, where we can see that the vanishing of the focus of attention trajectory is prevented by pumping the energy due to the inhibitory process. This is formally proven in the following analysis which strongly supports the need for the inhibitory process. The injected energy turns out to be particularly useful in the case of still images where, in absence of inhibition mechanism, are characterized by $\dot{\mu} = 0$ at each time step.

Theorem 2. *Suppose we are considering still images, i.e. $b(t, x) \equiv b(x)$ does not depend on the variable t. The attention trajectory does not vanishes only if there are regions $\mathcal{X} \subset \mathcal{R}$ such that for $\forall x \in \mathcal{X}$ we have $\dot{\mu}(x) \neq 0$.*

Proof. The proof is straightforward by contradiction. In case $\mu(x)$ is constant, the energy balance equation 4.13 yields

$$K(t) + U(t) + \lambda \int_0^t d\tau \dot{a}_i^2 = K(0) + U(0),$$

which is violated as $t \to \infty$ if the focus of attention trajectory does not vanish. □

We can promptly see that there should be at least a region $\mathcal{X} \subset \mathcal{R}$ such that $\forall x \in \mathcal{X}$ as $a(t)$ approaches x then $\log\|a - x\| < 0$ and, therefore, in order for M to become positive we need $\dot{\mu}(x) < 0$. This is in fact the outcome of the effect of the inhibitory function I which injects the energy M into the system.

4.6 Numerical issues

Let us consider the problem of simulating the dynamics of the system defined by equations (4.8), (4.9), and (4.10), subject to the boundary conditions $a(t_0) = a_0$, $\dot{a}(t_0) = \dot{a}_0$ and $I(t_0, x) = 0, \forall x \in \mathcal{R}$. In order to end up into the first-order canonical structure of differential equations let us introduce the auxiliary variable $z(t) = \dot{a}(t)$. Then the problem of determining $a(t)$ is equivalent to

$$\begin{cases} \dot{I}(t) = \beta \left(g\left(x - a(t)\right) - I(t) \right) \\ \dot{a}(t) = z(t) \\ \dot{z}(t) = -\lambda z(t) - (e * \mu)(t, a(t)). \end{cases} \tag{4.14}$$

Now, if we pose $y = (I, a, z)'$ and $u := (\mu_1, \mu_2)$ then system (4.14) can be compactly re-written in the canonical form

$$\dot{y} = \Phi(y, u). \tag{4.15}$$

that can be solved numerically by classic methods like Euler's and Runge-Kutta's.

We notice in passing that the last equation can be explicitly rewritten as

$$\dot{z}(t) = -\lambda z(t) + \frac{1}{2\pi}\int_{\mathcal{R}} dx \frac{a(t)-x}{\|a(t)-x\|^2}\big(\mu_1(x,t)(1-I(x,t)) + \mu_2(x,t)\big).$$

This means that for any pair $(a(t), I(x,t))$ the explicit computation of $\dot{z}(t)$ requires the correspondent numerical integration over the retina.

It is worth mentioning that the singularity of the integral for $x = a(t)$ requires an appropriate numerical treatment. This subject has been widely investigated in case of different improper integrals by using methods like change of variable, subtracting of singularity, and ignoring of singularity [67]. In this case we can naturally use the idea of subtracting of singularity, which simply consists of replacing the numerical contribution from the computation of the integral when approaching the singularity by an explicit computation. The following proposition suggests that if we assume that the virtual mass is nearly constant in a small box $\mathcal{B}$ of side ρ centered in $a(t)$ then we can remove the gravitational contribution from $\mathcal{B}$.

Proposition 1. *The following result holds for $i = 1, 2$:*

$$\int_{a_1-\rho}^{a_1+\rho}\int_{a_2-\rho}^{a_2+\rho} dx_1 dx_2 \frac{a_i - x_i}{(a_1-x_1)^2 + (a_2-x_2)^2} = 0. \tag{4.16}$$

Proof. Let us consider $i = 1$; the case $i = 2$ follows by symmetry. We have

$$\int_{a_2-\rho}^{a_2+\rho} dx_1 dx_2 \int_{a_1-\rho}^{a_1+\rho} \frac{a_1 - x_1}{(a_1-x_1)^2 + (a_2-x_2)^2} = -\frac{1}{2}\int_{a_2-\rho}^{a_2+\rho} dx_2 \log\Big((a_1-x_1)^2 + (a_2-x_2)^2\Big)\Big|_{a_1-\rho}^{a_1+\rho} = 0.$$

□

This proposition is clearly related to Gauss theorem and states a principle that is very well-known for 3D gravitational fields. No matter what is the space dimension, the reason for the null field in $a(t)$, whenever it is generated by any symmetric region (see e.g. Fig. 4.1-B), is basically related to the corresponding symmetry in the field that arises when considering the joint gravitational effect of masses $\mu(x - a(t))dx$ and $\mu(x + a(t))dx$. To sum up, from a numerical integration we can simply remove the pixel corresponding to the discrete position of $a(t)$ and use ordinary numerical methods for integration we get the best possible accuracy which corresponds with the image quantization.

4.7 Experiments

We carried out a massive experimental analysis to evaluate the quality of the trajectory of the focus of attention defined by $a(t)$, that comes out from the numerical integration of Eq. (4.15). The accuracy of the proposed model was tested in the scanpath simulation, as well as in the estimation of saliency maps. We remark that, unlike many related computational models, the proposed gravitational laws of focus of attention return a simulation of eye movements, and saliency maps are obtained as a by-product (Eq. (4.11)).

We experimented the model on a wide collection of datasets of static images and videos. We begin with the description of the dataset of human fixations used in the experiments. Data pre-processing is carefully described, as well as the procedure for the parameter estimation. Results on scanpath prediction and compare them with basic baselines and competitors. Finally, we present an accurate comparison with state-of-the-art models in saliency prediction.

Datasets

We selected 6 publicly available datasets to perform our experiments. We used 4 collections of images (MIT1003, SIENA12, TORONTO, KOOTSRA) and a video dataset (COUTROT) for scanpath prediction, while we followed a popular benchmark for saliency prediction (CAT 2000). In detail, we considered the following datasets:

- **MIT1003** [42]. This dataset contains 1003 natural indoor and outdoor scenes. They are sampled with variable sizes, where each dimension is in 405-1024px. The database contains 779 landscape images and 228 portrait images. Fixations of 15 human subjects are provided for 3 seconds of free-viewing observation.

- **SIENA12** [81]. Twelve grayscales images are chosen with the purpose of minimizing the semantic content. Stimuli include natural scenes, human constructions, but also abstract contents. The resolution of the images is 1024x768px. Fixations of 23 human subjects are provided for 5 seconds of free-viewing observation.

- **TORONTO** [11]. A collection of 120 color images of outdoor and indoor scenes. Resolutions of the images is 681x511px. Fixations of 20 human subjects are provided for 4 seconds of free-viewing exploration. A large portion of images do not contain specific regions of interest.

- **KOOTSTRA** [46]. This dataset includes 99 color images with symmetrical natural objects, animals in a natural contest, street scenes, buildings and natural landscapes. Resolution of the images is 1024x768px. Fixations of 31 human

subjects are provided for 5 seconds of free-viewing exploration. The category of the natural landscapes is more represented than the others.

- **COUTROT DATASET 1** [18]. It is a collection of 60 video clips. Categories include one or several moving objects, landscapes and scenes of people having a conversation. Resolution of the frames is 720x576 px. Fixations of 72 human subjects are provided. The average duration of each clip is 17 seconds. Some videos include ego-motion.

- **CAT 2000** [8]. We select the publicly available portion of this dataset, that contains 2000 images from 20 different categories. Stimuli include basic features (basic patterns, sketches, fractals), noisy and low resolution images, natural landscapes, abstract pictures (cartoon and line drawing), high level semantic contents (social, affective, indoor), and more. The resolution of the images is 1920x1080 px. Saliency maps are provided for each image. Maps are calculated with data on 18 different human subject free-viewing exploration.

Gathering data related to human fixations on visual stimuli (needed for the task of scanpath prediction) is a very expensive procedure. For this reason, some datasets are quite small. In order to make our analysis on scanpath prediction more robust, we merged the data from the aforementioned 4 collections (MIT1003, SIENA12, TORONTO, KOOTSRA), thus generating a unique, larger set called *FixaTons*, and correlated it with a software library for data usage and metrics computation. More details about this project advised by MIT Saliency Team are given in the appendix B.

Data pre-processing

All images and video frames are resized to a resolution of 224×224 pixels and converted to grayscale. The spatial gradient of brightness and optical flow are calculated with standard functions of the OpenCV library, that are related to the classic method described in [35] by Berthold K.P. Horn and Brian G. Schunck. The integration of equation (4.15) that drives the focus of attention trajectory is based on the `odeint` function of Python SciPy library. The function is based on LSODA, which is a general purpose software that dynamically determines where the problem is stiff and chooses the appropriate solution method [58]. The G-EYMOL model that we propose in this chapter is generic, and it allows us to use exactly the same experimental setup for both static images and video: image are just regarded as videos whose frames are identical at each time step.

The scenes we watch every day are often affected by camera motion due to the cameraman activity, as well as the vibrations of the camera itself. Natural clips can include many moving objects at different depths and speeds, so that scenes get to be extremely chaotic. The treatment of such complex scenes goes beyond the scope

of this paper. Further work will have to include more solid estimates of pixel velocities: methods for camera motion estimation [3, 24, 34] and background subtraction [23, 84] are already present in the literature and can help in creating more suitable estimates for the proposed model.

Parameters estimation

While, for the sake of simplicity, we chose to associate the focus of attention $a(t)$ with the unitary mass, different choices of the parameters of the system can bring very different behaviours. In particular, the weights α_1 and α_2 in equations 4.1 and 4.2 can emphasize the contribution of the two different features and determine what is more relevant and, hence, they value has to be selected. In our experiments, we started by setting α_1 and α_2 to initial values that we found to be qualitatively satisfactory when running our algorithm on the video stream coming from cameras of our laboratory or some sample natural videos. Then, we measured the performance on the task of saliency prediction as a criterion that we aim at maximizing by a greeds search in the parameter space. In particular, starting from the initial α_1 and α_2, we search for improvements in the saliency prediction score (AUC-Judd metric [42]) by considering the following four pairs of values

$$(\alpha_1 + \delta_{\alpha_1}, \alpha_2),$$

$$(\alpha_1 - \delta_{\alpha_1}, \alpha_2),$$

$$(\alpha_1, \alpha_2 + \delta_{\alpha_2}),$$

$$(\alpha_1, \alpha_2 - \delta_{\alpha_2}),$$

where $\delta_* = 0.01$. We repeat the search procedure until the method converges to the best configuration or the maximum number of iterations is reached. The algorithm is described in more details in the Algorithm 2.

The MIT1003 data and the COUTROT videos, that we use for scanpath prediction experiments, also come with the saliency maps, and we exploited them to estimate the model parameters. In particular, we used COUTROT and a randomly selected subportion of 100 images from MIT1003. We remark that α_2 is about the pixel velocities, so we mostly focus on video data. Since the described search is only locally optimal, we repeated the entire procedure for n times, using different random data subsamples of MIT1003 ($n = 30$). The overall best pair (α_1, α_2) is selected and used for the following experiments.

The positive parameter λ in equation 4.10 is more strongly connected to the system dynamics, as it determines how quickly the oscillations are damped to converge to a precise target. Similar considerations apply to the parameter β in the definition of the function of Inhibition-of-Return in equation 4.8, which determines how

Algorithm 2 This procedure describe the search algorithm for the optimization of the model parameters. The function `score()` correspond to the AUC-Judd metric calculated on a random sample of 100 input stimuli from MIT1003 and on the COUTROT dataset.

```
1: procedure ParametersOptimization
2:     Select an initial state i ∈ D
3:     maxIter ← 100
4:     numIter ← 0
5:     do
6:         isImproved ← False
7:         for j ∈ Neighbour(i) do
8:             if score(i) < score(j) then
9:                 i ← j
10:                isImproved ← True
11:            end if
12:        end for
13:        numIter ← numIter + 1
14:    while isImproved and numIter < maxIter
15: end procedure
```

quickly the masses are suppressed and attention is shifted to targets far from the currently attended location. We chose these quantities to satisfy the criterion of approximating human behaviour under certain global statistics of eye movements. In particular, given the same data and the same greedy search used to validate α_1 and α_2, we searched for (λ, β) that guarantees a fixation rate (whose computation is described below) that is as more similar as possible to the human fixation rate (3.5 fixation per second).

Scanpath prediction

Most state-of-the-art computational models in visual attention estimate the probability distribution of fixating a certain image location, i.e. the saliency map [11, 17, 28, 38, 42, 49] and do not produce a temporal sequence of eye movements (sequence of fixations), which can be of great importance for understanding human vision as well as for building systems that deal real-time with video streams or need of a meaningful visual exploration process [77]. In [7] many saliency models have been evaluated in the task of scanpath prediction.

In this section, we compare the scanpath prediction performance of our model with two baselines (Random and Center [12]), a saliency map based model (Itti's model [38]), and with the simple model EYMOL [80]:

- *Random.* Fixations are sampled from a uniform distribution.

- *Center.* Fixations are sampled from a Gaussian distribution, centered in the middle of the image. The center baseline is a very good saliency predictor [41, 42]. The majority of human fixations appear to be next to the center. This is due to a viewing strategy by which subjects first inspect the image center, probably to rapidly gather a global view of the scene [70] or because of the photographer bias to put interesting object in the middle of the scene [7, 8].
- *Itti.* Fixations are generated by the procedure described in the original paper [38]. A saliency map is calculated in advance, then the algorithm described in [44] is used to generate fixations. Fixation locations are selected by a *Winner-Take-All* mechanism, while the *Inhibition-Of-Return* mechanism suppresses activity in the selected location and leads to different locations. Although this approach has biological argumentation [44], it suffers from several problems: many calculation steps needed for each fixation, it performs global calculations across the field of view, and it is unclear how to extend it to the case of video streams.
- *Eymol.* A recently proposed dynamic model of focus of attention [80]. We followed the same strategy of [80] to generate trajectories of visual exploration.

Scanpath metrics. In order to evaluate the behavioural properties of our model, of the two baselines, and of the competitors, in generating simulated scanpaths, we measure the similarity to human scanpaths. In the neuroscience literature, the two main metrics proposed to measure the distance/similarity between two sequences of fixations are:

- *"String-edit" or Levenshtein distance* (*distance*). The input stimulus (the input image) is divided into $m \times m$ regions, labeled with characters. Scanpaths are turned into strings by associating each fixation with the character of the corresponding region. Finally, the string-edit algorithm is used to measure the distance between the two generated strings ($m = 5$). The dynamic program to compute this metric is described in [43]. The metric has been used in different works for comparing different human scanpaths [10, 26]. A similar version of this metric is also used in [7] to evaluate computational models of saliency in the task of scanpath prediction. It has been shown [15] that this metric is robust to changes in the number of regions used to divide the input stimulus.
- *Scaled time-delay embedding* (*similarity*). Time-delay embedding similarity is commonly used in order to quantitatively compare stochastic and dynamic scanpaths of varied lengths. This similarity is popular in dynamic system analysis and carefully described in [75]. In particular, let us consider the problem of comparing a *simulated* (s) and a *human* (h) scanpath,

$$s \equiv s(1), ..., s(a)$$

$$h \equiv h(1), ..., h(b)$$

eventually of different length, $a \neq b$. We indicate with $\mathcal{C}_s^k(t) = (s(t), ..., s(t+k))$ a k-dimensional sub-sequence of fixations extracted from s, starting from the t-th fixation, and we indicate with

$$X = \{\mathcal{C}_s^k(t)\}_{t \in (1,...,a-k)} \subset \mathbb{R}^k \tag{4.17}$$

the space of all this possible k-dimensional sub-sequences extracted from s. Analogously,

$$Y = \{\mathcal{C}_h^k(t)\}_{t \in (1,...,b-k)} \subset \mathbb{R}^k \tag{4.18}$$

is the the space of all this possible k-dimensional sub-sequences extracted from h. Notice that $k < \min\{a, b\}$. Comparison between the clouds X and Y of data points in $\mathbb{R}^k$ will reflect dynamical similarities between the two scanpaths. The time-delay embedding distance $\text{tde}_k(\cdot, \cdot)$ is defined as

$$\text{tde}_k(s, h) = \frac{1}{|X|} \sum_{x \in X} d_k(x, Y) \tag{4.19}$$

where

$$d_k(x, Y) = \min_{y \in Y} \{\|x - y\|\}, \tag{4.20}$$

i.e., tde_k it is the *mean minimal* distance. In our experiment we used a scaled version [81], where fixations coordinates are normalized in $[0, 1]^2$ to deal with the fact that images from the considered datasets may significantly differ in resolution. In order to generate a final score that evaluates all the possible values of k, we propose the scaled time-delay embedding similarity to be defined as

$$\text{stde}(s, h) = \exp^{-\frac{1}{|K|} \sum_{k \in K} \text{tde}_k(s,h)}, \tag{4.21}$$

where $K = \{1, ..., \min\{a, b\} - 1\}$ is the set of all possible sub-sequences lengths. Notice that $\text{stde}(s, h) = 1$ indicates perfect similarity, while as it approaches zero the more dissimilar the scanpaths are.

Results. Since the output of the proposed model produces a simulated continuous trajectory, it is necessary to extract fixations from such trajectory in order to compute the just described metrics. This is commonly done with human data with algorithms that use thresholds on the distance of the sampled gaze positions and on the time spent in certain coordinates. We followed the approach of [19], implemented in the standard Python library PYGAZE, that extracts fixations from raw data of an eye-tracker device. Threshold values have been set with default values, that are designed to extract only human-like fixations. The same method is also exploited in the case of the competitor Eymol [80].

Tables 4.1 and 4.2 summarize experimental results on scanpath prediction, distinguishing between image and video data. In the case of string-edit distance smaller values indicate better performances, while in the case of scaled time-delay embedding similarity larger values are preferable. For each compared model, the same procedure is used to compute the two scores (string-edit distance and scaled time-delay embedding similarity). In detail, for each input stimulus (image or video) we are given a set of human scanpaths (from 15 to 72). For each human scanpath, a simulated scanpath of the same length is generated by the considered model, and the two matching scores are computed. Then, we focus on two quantities, that are the mean of the matching scores for the given stimulus (MEAN) and the score associated to the best predicted human scanpath (BEST). Finally, we report the mean and standard deviations of such quantities over the entire data collection.

The proposed model, G-EYMOL, achieves better results in each of the analyzed cases against the two proposed baselines and the competitors. This confirms that approaching the scanpath prediction problem by modeling the focus of attention as a dynamic process is a promising direction, as already introduced by EYMOL in [80], and that the gravitational-inspired solution of G-EYMOL more closely resembles the human behaviour. In order to better grasp the improvement introduced by G-EYMOL, in Table 4.5, we report the cumulative score curves in the setting of Table 4.1. The proposed model allows us to get a larger fraction of matches that are associated to small string-edit distances and to larger time-delay similarities. In the case of the COUTROT video dataset, the human scanpaths are longer then scanpaths on images. This is because the average duration of the videos is 17 seconds, while the images were observed from 3 to 5 seconds each. This explains why the values of the string-edit distance are different between the two cases. In contrast, scaled time-delay embedding similarity is less sensitive to scanpath length variations. *Itti* is not included in the comparison for scanpath estimation in the case of videos since the algorithm is defined by the authors only for the case of static images.

In several real world simulations, the proposed model shows interesting behavioural properties, see Figure 4.4 for some examples or `https://sailab.diism.unisi.it/visual-attention-modeling/` for other simulations and an online demo (we also publish the link to the code repository of our model). For a naive observer, simulated scanpaths are difficult to distinguish from human scanpaths. In the simulations reported in the project website, we also naively distinguish among fixations and rapid movements between consecutive fixations (saccades) using a red or blue markers, respectively, and it is easy to see the emergence of the tracking behaviour for salient moving objects.

Model	**DATASET COLLECTION** [11, 42, 46, 81]			
	String-Edit (distance)		**Scaled Time-delay embedding (similarity)**	
	Mean	**Best**	**Mean**	**Best**
G-Eymol	**7.34 (2.42)**	**3.72 (1.92)**	**0.81 (0.03)**	**0.85 (0.03)**
Eymol	7.94 (2.46)	4.10 (1.95)	0.74 (0.07)	0.81 (0.07)
Itti	8.15 (2.48)	4.36 (1.94)	0.70 (0.09)	0.76 (0.09)
Center	8.13 (2.42)	4.35 (1.90)	0.72 (0.04)	0.77 (0.04)
Random	8.21 (2.40)	4.43 (1.87)	0.70 (0.04)	0.75 (0.04)

Table 4.1: **Results on scanpath prediction (Data collection).** Results on a collection of four image datasets: MIT1003 [42], SIENA12 [81], TORONTO [11], KOOTSTRA [46]. For each stimulus (image), the dataset has a set of human scanpaths of variable cardinality. For each stimulus, we calculate the metrics for each of these human scanpaths. The MEAN score is averaged over each stimulus, while BEST is the score of the best prediction for the considered stimulus. The table reports mean and standard deviation (in brackets) of these scores for the entire data collection.

Model	**COUTROT DATASET** [18]			
	String-Edit (distance)		**Scaled Time-delay embedding (similarity)**	
	Mean	**Best**	**Mean**	**Best**
G-Eymol	**35.68** (13.97)	**23.83** (13.07)	**0.79** (0.05)	**0.86** (0.04)
Eymol	39.90 (11.29)	30.48 (10.76)	0.77 (0.03)	0.84 (0.03)
Center	44.24 (2.24)	36.68 (1.41)	0.74 (0.01)	0.79 (0.001)
Random	45.51 (2.97)	38.45 (1.33)	0.70 (0.01)	0.76 (0.01)

Table 4.2: **Results on scanpath prediciton on videos (COUTROT)** Results on the video dataset COUTROT [18].For each stimulus (video), the dataset has a set of human scanpaths of variable cardinality. For each stimulus, we calculate the metrics for each of these human scanpaths. The MEAN score is averaged over each stimulus, while BEST is the score of the best prediction for the considered stimulus. The table reports mean and standard deviation (in brackets) of these scores for the entire data collection.

Figure 4.4: **Simulated scanpaths with G-EYMOL.** This figure shows some outputs of our model in a task of free-viewing of sample stimuli from the dataset MIT1003 [42]. The blue square indicates the stating point of the scanpath. Larger arrows are associated to longer transitions. We can observe that small or big objects as well as faces attract attention. This is certainly due to the fact that they present high values of brightness gradient at the contours. Notice how the inhibition of return mechanism allows wide exploration of the scenes that guarantees a good acquisition of the information.

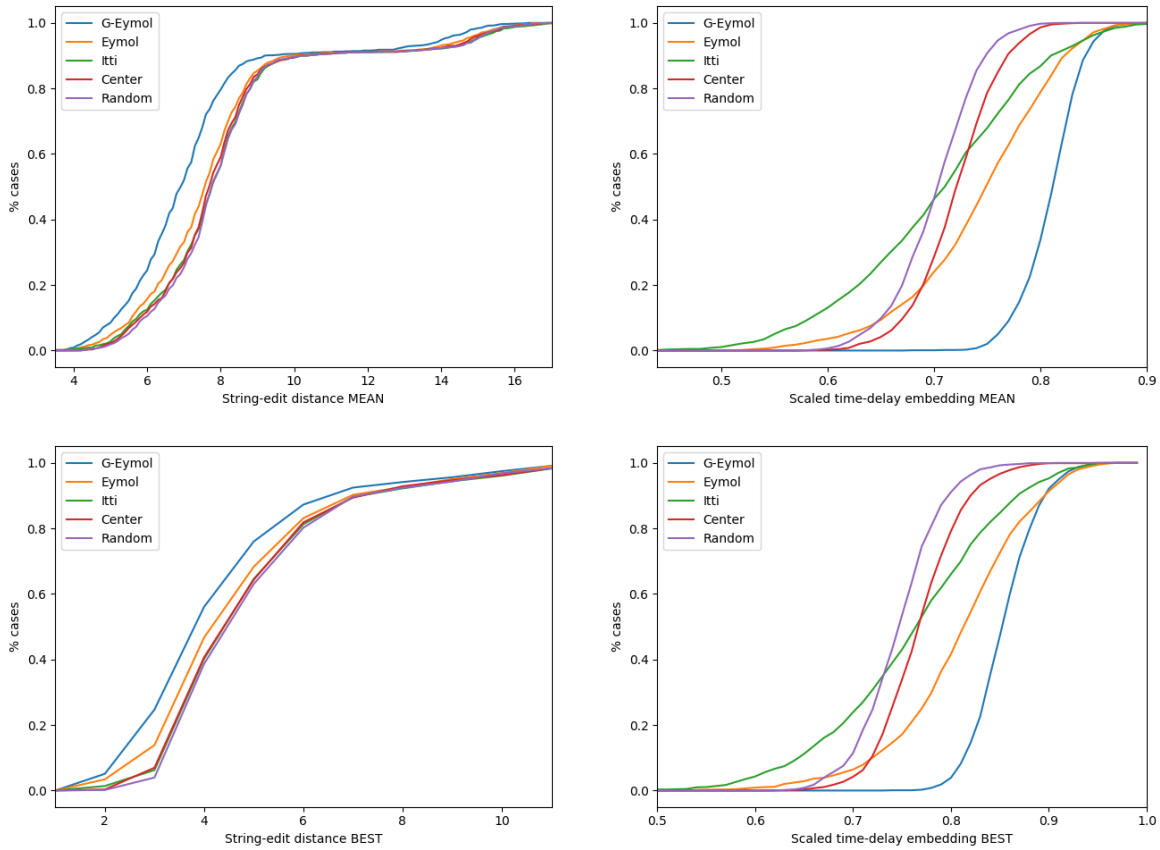


Figure 4.5: **Cumulative score curves in scanpath prediction.** For each value of the string-edit distance (left) and of the scaled time-delay embedding (right), we report the percentage of input stimuli (i.e, the percentage of images in the setting of Table 4.1) for which a given model obtains a score less than or equal to that value.

Saliency prediction

Several models have been proposed by the computer vision community to address the problem of predicting saliency maps. They usually differ in the definition of saliency. For instance, it has been claimed that the attention is driven according to a principle of information maximization [11] or by an opportune selection of surprising regions [36]. Machine learning approaches have been used to learn saliency directly from data. Judd *et al.* [42] collected 1003 images observed by 15 subjects and trained an SVM classifier with low-, middle-, and high-level features. Current top level performance is achieved by machine learning methods[17, 49, 74]. A detailed description of the state of the art is given in [7]. All of this methodologies postulate a central role of the saliency map.

In contrast, our model describes scanpath of free visual exploration of images or videos and saliency maps can be obtained as a by-product (see equation 4.11). We compare our model with state-of-the-art models in the CAT2000 benchmark. This is the largest dataset of saliency maps obtained from human fixations, and both continuous saliency maps and discrete fixation locations are provided. It is composed by several different semantic categories, which makes the evaluation robust.

Different scanpaths are obtained by varying initial condition of the differential system. It has been argued that models that have some blur and center bias on the saliency map can improve their score with respect to some metrics [13]. For this reason, we collected data of 30 independent trials[2], each run was randomly initialized almost at the center of the image and with a small random velocity, and integrated for a running time corresponding to 5 second of visual exploration. The result was blurred with Gaussian blur, and saliency map is then calculated by summing up the most visited locations. A grid search over *blur radius* and *center bias* parameters have been used, in order to maximize AUC-Judd [42] and NSS score on the data of CAT2000. In AUC-Judd the saliency map is considered as the output of a binary classifier of the fixations, then the Area Under the Curve (AUC) score is computed. NSS is the Normalized Scanpath Saliency between two saliency maps, and it is measured as the mean value for the normalized saliency map at fixation locations.

Results. Table 4.3 shows the performance of the proposed model compared to other state-of-the-art models. Despite the extreme simplicity of the feature used with in the proposed model, it outperforms all unsupervised models [11, 28, 38] with respect to AUC-Judd metric, and it gets results not far from the supervised approaches. In the case of NSS, our model shows a result that is slightly below the one of EYMOL [80]. This is due to the fact that EYMOL can produce maps that are more biased toward the center of the image, since it does not implement any Inhibition of Return mechanisms, and the NSS metric is strongly influenced by the center bias

[2]We observed that more trials would not produce improvements in the performance.

[7]. We also remark that the scanpaths produced by EYMOL are less human-like than the case of G-EYMOL, as shown in the previous experiments of scanpath prediction. This suggests that even if both EYMOL and G-EYMOL generated similar saliency maps, they would be the outcome of strongly different scanpaths, and the case of G-EYMOL would be preferable. It is also worth mentioning that our model operates with a single basic feature such as the spatial gradient, since the optical flow is identically zero on each sample image, while the other models can count on a large number of hand-crafted features [42] or rich feature representation extracted from deep learning models [17, 49, 74] and fully supervised data. In addition, all supervised competitors model have been designed specifically to estimate the saliency map and none of them produce eye movements.

Finally, in Figure 4.6 we report some qualitative examples of saliency maps generated by our model.

Model	Reference	Supervised	CAT2000	
			AUC	NSS
G-Eymol	our	No	**0.836** (0.001)	1.57 (0.04)
Eymol	our	No	0.83	**1.78**
AIM	[11]	No	0.76	0.89
Itti-Koch	[33, 38]	No	0.77	1.06
AWS	[28]	No	0.76	1.09
Judd Model	[42]	Yes	0.84	1.30
eDN	[74]	Yes	0.85	1.30
DeepFix	[49]	Yes	0.87	2.28
SAM	[17]	Yes	0.88	2.38

Table 4.3: **Results on saliency prediction (CAT2000).** Comparison with state-of-the-art models on the benchmark of saliency prediction. We also report the results of fully supervised models.

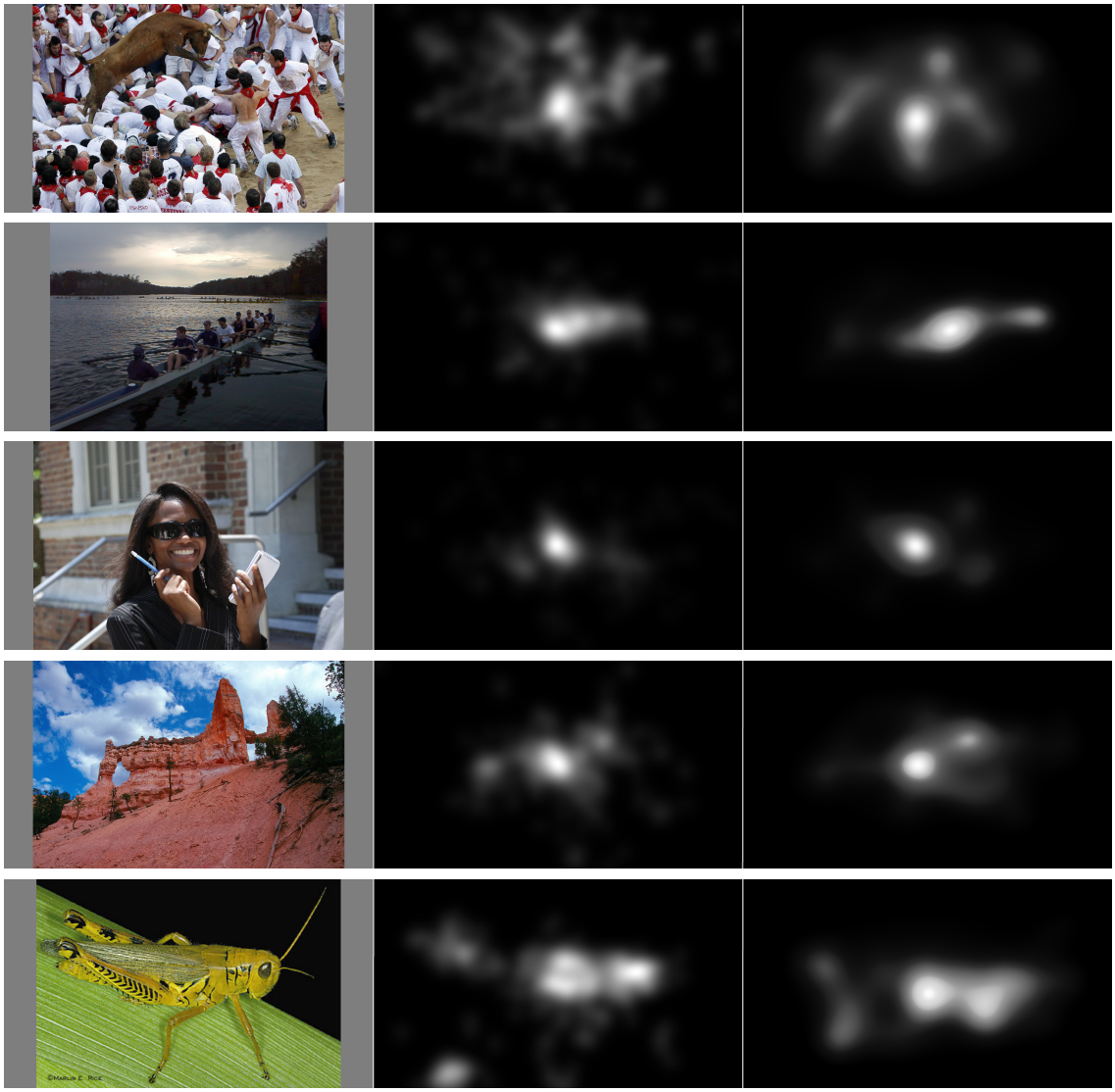


Figure 4.6: **Saliency map with G-EYMOL.** Each row present in order the input stimuli (first column), human saliency map (second column) and the saliency map predicted with our model (third column).

4.8 Discussion

In this chapter, we have introduced G-EYMOL, a computational model of visual attention where the focus of attention is subject to a gravitational field. The distributed virtual mass that controls eye movements is associated with the presence of details and with motion in the video. When restricting to motion and details, we are basically modeling attention at early stage of vision, which somewhat corresponds with involving feature maps of V1 zone of the human visual system. In this area, in addition to the information about visual details, a raw velocity tag is attached to every location of the visual field [21, 68, 73]. Interestingly, the inhibition of return mechanisms that avoid to get stuck in high saliency regions are naturally carried out by an appropriate modulation of the associated virtual mass, which leads to an overall dynamic model that very well matches human behaviour.

The definition of saliency provided in equation 4.11 collapse to the classical definition for $\theta = 0$; however it becomes interesting in the case θ grows since it leads to values of the saliency which emphasize the recent behaviour of the trajectory. This can be seen as a dynamic definition of saliency and it is worth to conduct future investigation about the effectiveness of this formulation.

A distinctive feature of the proposed approach is that the attraction of the focus of attention turns out to be a unique process, regardless of the stimulating source of attention. While our study has been restricted to stimuli based on gradient of the brightness and on motion, the extension of the theory to the case in which the focus of attention trajectory is additionally stimulated by a field extracted from semantic-based features seems very promising. As pointed out in the recent work in [4], the training of convolutional networks benefits from processing information that is selected by the focus of attention. Hence, it can gain a circular reinforcement, thus exploiting refined valued of the developed visual features. Top level performances are achieved especially in the prediction of scanpath, which is the primary purpose of the proposed computational model.

Chapter 5

Conclusions

In this thesis, we defined three different models of visual attention. Their performance have been evaluated using a corpus of images and videos and eye-tracking data collected experimentally by different international renewed institutions. Parts of this data has also been collected by our group. The three models are based on fundamental laws and principles of mechanics. The principal intent is to generate sequences of region of interest predicting human eye fixations, based on visual low level properties of the image or video.

In one case, described in chapter 3, a convolutional neural network is employed.This case is a first attempt to integrate information from outside. This is a promising feature. It allows the attention system to be integrated with vision models. If these vision models are task oriented, scanpaths will also be functional to the task. In the experiments shown in this chapter is observed in a qualitative way as the scanpaths are modified and directed in an evident way towards the main objects of the scene. This is due to the fact that a priority map from a fully-convolutional network trained for object recognition has been integrated. Future investigations are necessary. Although, as discussed, this seems a happy computational instance of the model hypothesized by Yarbus [79], it is necessary to verify the behaviour on a large set of different tasks. It also seems promising to see how this can fit into the Curiosity Driven Learning.

The possible applications of these models are many. Vehicles travelling in hostile environments, exploration of space and satellite images, or submarine scenarios can benefit from a model that quickly focuses points of interest of a scanpath on which to perform subsequent and more accurate analysis. The models can also be used for the study of degenerative diseases or mental disorders. Having a sufficiently large database of data of healthy and sick subjects, it would be possible to search in the space of the parameters of the model those that most characterize certain groups of subjects. This would have a double usefulness, analytical and diagnostic.

The reader may wonder which are the cases in which it is preferable to use one

of the three proposed models rather than another. We tried to highlight the differences during the work whenever possible. Here we will try to summarize the main differences. Even if no explicit experiments have been made to evaluate the execution times, it can be said that EYMOL is the cheapest of the three models from the point of view of the amount of calculations required. Compared to CF-EYMOL it obviously doesn't need to generate the activation map with the inception-v3 model. Moreover, the calculation of functional terms for EYMOL is local and involves, for each time instant, a limited number of pixels around the fixation point. G-EYMOL, on the contrary, requires the computation of a term which involves a sum over the whole retina to get the resulting contribution. EYMOL is therefore to be preferred in those cases in which the computational cost is particularly relevant. For the task of saliency estimation, the difference in performance between the three models is small. One might still prefer to use EYMOL for a simple and fast saliency predictor, or use CF-EYMOL to get the best estimation. However, it is worth noting that if only a saliency estimation is required without the need of an online exploratory process, other models of literature are preferable. They are faster because they do not need to simulate numerous observations [11, 38], or they are more accurate in estimating salience [17]. The scanpath simulation, on the other hand, sees G-EYMOL in clear advantage over the other two approaches and all competitors. The introduction of the sum over the entire retina allows to spot stimuli which are far from the actual point of fixation very quickly. The motion feature is also very relevant in the case of video analysis. Finally, the introduction of the Inhibition of Return mechanism makes the scanpaths very similar to human ones. The G-EYMOL model is therefore to be preferred to simulate processes of image exploration and in which the order of exploration is important. Its great similarity with human eye movements makes it suitable for applications in robotics, for example, to improve the interaction between humans and humanoid.

The proposed models are interesting not only under the perspectives of application in the field of artificial intelligence, but also for the comprehension of the human vision itself. While the mechanics by which eye movements take place are very much studied, we are still far from a unifying theory that jointly explains how these depend on visual input and on an internal state of the individual (or goal). Subsequent studies are desirable that put the model even more closely in relation to human vision processes. Several points need further study. The top-down component has not been described in this work, except with slight hints (see chapter 3). While it has been suggested that feature maps come from learning systems and indicate to a certain extent a preference index of each location, and that this process can be used to induce task-oriented ocular movements, a large experimentation in this direction is necessary. Another promising and potentially complementary study is to understand the mechanisms that involve the variation of pupil diameter. A lot

of data exists and could be exploited to refine the model or automatic processes of information acquisition. The model that generates the eye movements should be closely linked to this process of information acquisition, since it is clear that the two are directly dependent on each other. Finally, another highly variable and informative parameter in humans is the fixation duration. For now, an artificial "clock" has been inserted in the different models (EYMOL and CF-EYMOL) or this duration has been defined together with the inhibition on return (G-EYMOL). In both cases, the fixation duration is fixed a priori and does not depend in any way on the task. A description of how the task influences this parameter in humans could inspire algorithms for machine vision that increasingly emulate and approximate human ability, and move a bit forward in the understanding of the complex and refined process of vision so perfectly designed by the nature.

Appendix A

The Least Action Principle

A.1 The Least Action Principle

For any physical system, the Lagrangian is defined as the kinetic energy less the potential energy. In symbols,

$$L = K - U, \tag{A.1}$$

where K is the kinetic energy and U is the potential energy. The action is defined as the integral between two time instants t_1 and t_2 of the Lagrangian, i.e.

$$\mathcal{S} = \int_{t_1}^{t_2} L(t, x, \dot{x})\, dt \tag{A.2}$$

The Principle of Least Action states that, *in any physical system, the path an object actually takes minimizes the action* [29] [25]. It can be shown that extrema of an action occurs at

$$\delta\mathcal{S} = 0, \tag{A.3}$$

that is true for $x(t)$ which satisfy

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = \frac{\partial L}{\partial x}. \tag{A.4}$$

The last is called Euler-Lagrange equation. But notice, Euler-Lagrange equation is only a necessary condition for a minimum.

Let us suppose an object m, within a potential field $U(t, x)$, starts its true path $x(t)$ from a position $x_1 = x(t_1)$ and ends at position $x_2 = x(t_2)$. If we calculate for each time instant the difference between its kinetic energy and its potential energy, and then integrate this with respect to time from t_1 to t_2, in formulas

$$\mathcal{S}_{[t_1,t_2]}[x] = \int_{t_1}^{t_2} \frac{1}{2}m\dot{x}^2 - U(t, x)\, dt, \tag{A.5}$$

we would find that *this integral is least along the true path, for close enough* t_1 *and* t_2. That is, if we calculate the same quantity for any other path, say $\bar{x}(t)$, starting from $x_1 = x(t_1)$ and ending in $x_2 = x(t_2)$, this would be bigger. In formulas, we could write

$$\mathcal{S}_{[t_1,t_2]}[x] < \mathcal{S}_{[t_1,t_2]}[\bar{x}]. \tag{A.6}$$

We now compute the first variation, that is the variation on the first order between the action calculated for the true path $x(t)$ and the action calculated for a nearby path

$$\bar{x}(t) = x(t) + h(t).$$

We have that $h(t)$ can be any function, but the idea is to chose it very small. Also, since both $x(t)$ and $\bar{x}(t)$ start at x_1 and end at x_2, we do not let it vary at the extrema, that is

$$h(t_1) = h(t_2) = 0. \tag{A.7}$$

The action calculated for the path $\bar{x} = x + h$ (omitting time dependences for a while) is

$$\mathcal{S}_{[t_1,t_2]}[\bar{x}] = \int_{t_1}^{t_2} \frac{1}{2} m \left(\dot{x} + \dot{h}\right)^2 - U(t, x + h)\, dt \tag{A.8}$$

If we consider h very small, and eliminating second and higher order terms, we have that

$$U(t, x + h) \approx U(t, x) + h \cdot U'(t, x). \tag{A.9}$$

About the kinetic energy we have that

$$\frac{1}{2} m \left(\dot{x} + \dot{h}\right)^2 = \frac{1}{2} m \left(\dot{x}^2 + 2\dot{x}\dot{h} + \dot{h}^2\right) \approx \frac{1}{2} m \left(\dot{x}^2 + 2\dot{x}\dot{h}\right), \tag{A.10}$$

also in this case omitting second and higher order terms. We can hence re-write (A.8) in the form

$$\mathcal{S}_{[t_1,t_2]}[\bar{x}] = \int_{t_1}^{t_2} \frac{1}{2} m \left(\dot{x}^2 + 2\dot{x}\dot{h}\right) - U(t, x) + h \cdot U'(t, x)\, dt. \tag{A.11}$$

It is of immediate verification that the variation to the first order coincides with

$$\delta\mathcal{S} = \int_{t_1}^{t_2} m\dot{x}\dot{h} + h \cdot U'(t, x)\, dt. \tag{A.12}$$

Integrating by parts the first term, we get

$$\delta\mathcal{S} = m\dot{x}h\big|_{t_1}^{t_2} - \int_{t_1}^{t_2} \frac{d}{dt}\left(m\dot{x}\right) h + h \cdot U'(t, x)\, dt. \tag{A.13}$$

The integrated part disappears thanks to the condition A.7, and the rest can be grouped in this way

$$\delta\mathcal{S} = -\int_{t_1}^{t_2} \left(m\ddot{x} + \cdot U'(t, x)\right) h\, dt. \tag{A.14}$$

If we are not lost at this point, we remember that the Principle of Least Action states that the first variation $\delta\mathcal{S}$ vanishes along the true path. Since it must happen whatever $h(t)$ is, the only way to the integral (A.8) to vanish is the path to satisfies

$$m\ddot{x} + \cdot U'(t, x) = 0. \tag{A.15}$$

The last correspond to the Euler-Lagrange equation calculated on the action. Observe that it is just the Newton law of dynamics

$$F = ma \tag{A.16}$$

for a gravitational potential field, or the harmonic oscillator equation

$$\ddot{x} = -kx \tag{A.17}$$

using the appropriate potential.

As an important remark, please notice that we needed to have *given* initial and final position of the object m in our computation. Precisely, it allowed us to eliminate integrated term, after integration by parts. In fact, in order to apply the Principle of Least Action right in the same way we have done, we must know something about the future.

Determinism and Least Action

There is an assumption that is tacitly accepted while deriving laws of nature with the Principle of Least Action. For a given initial condition, the future of the system is uniquely determined. It is called determinism, or Laplacean Demon. It does not work in general, but it does hold for mechanical systems. If initial position and velocity are known at time t_1, there exists a unique possible final state at t_2.

In our cases of study, initial conditions of the system are given

$$x(t_1) = x_1$$

$$\dot{x}(t_1) = v_1$$

together with the potential energy

$$U(t, x(t)).$$

Differently from the case in which we enunciated Principle of Stationary Action, here we have given initial position and initial velocity instead of initial and final position. We know nothing about the future states. Extremal are not fixed and we actually need laws to predict the final position. But even if we do not explicitly know the final position of our trajectory, we assume it has one and only one $x(t_2) = x_2$.

That is the determinism assumption. We do not know it explicitly, but we know it exists and is unique. Then, admissible paths cannot vary this final position. This means we can step-by-step follow the algorithms of the previous section and, by the principle of stationarity of the action, get the differential laws which describe the dynamics of the mechanical system we are modeling. Moreover, if L, $\frac{\partial L}{\partial x}$ and $\frac{\partial L}{\partial \dot{x}}$ are continuous in $[t_1, t_2] \times \mathbb{R}^2$, there a nice property holds, that is that the solution of (A.4) is unique (Cauchy theorem for second order IVP).

A.2 An example: the harmonic oscillator

Let us consider the unidimensional harmonic oscillator. In classical mechanics, it is a system that experiences a resisting force proportional with the displacement from the equilibrium position. In formulas, the system dynamics are governed by the differential law

$$F = -kx, \tag{A.18}$$

where k is a constant. Analogously, by the Principle of Stationary Action we can enunciate the integral counterpart of the law. The true path of an object m in an harmonic oscillator system, with initial position $x(t_1) = x_1$ and final position $x(t_2) = x_2$, is the one for which action is stationary with respect to nearby paths. In formulas, if we define the action

$$\mathcal{S}\left[x(t)\right] = \int_{t_1}^{t_2} \frac{1}{2} m\dot{x}^2 - \frac{1}{2} kx^2 \, dt, \tag{A.19}$$

then the true path is a path along which the first variation is zero. The true path is then

$$x(t) \in C^\infty \text{ such that } \delta\mathcal{S}\left[x(t)\right] = 0. \tag{A.20}$$

As we have seen, it corresponds to find the solution of the Euler-Lagrange equation. In this case, Euler-Lagrange equation is

$$m\ddot{x} = -kx \tag{A.21}$$

which exactly correspond to (A.18). Notice, we derived the differential form of the law. But while (A.18) tells a local information, with (A.20) we have the further information that - using the same words of Feynman - the object *smells the neighboring paths to find out whether or not they have more action*[1], and then it choses the least.

[1]Some very interesting discussions on the fact that the object *really smells* all possible paths can be found in [25]

Appendix B

FixaTons

B.1 Overview

Supervised data is often very expansive to collect. This is particularly true in the case of human fixations data where the need of eye-tracking device and guarantees of correct environment conditions do not even allow remote collaboration and extensive crowd-sourcing.

Some alternatives to eye-trackers have been proposed. A mouse-contingent multi-resolutional paradigm [39] based on neurophysiological and psychophysical studies of peripheral vision, to simulate the natural viewing behaviour of humans using common mouse instead of an eye tracker to record viewing behaviours, thus enabling large-scale crowd-sourcing. While paradigms like the one just described can are validated and can help collection big amount of human saliency data, still real gaze information coming from eye-trackers is necessary for guarantees of good quality information as well as for studying actual dynamics of eye movements and visual attentive behavioural statistics like saccade velocity, fixation duration, fixations per second, and more.

A certain number of datasets of human behavioural data about free visual exploration is publicly available. However, this datasets are often small, task specific, semantically biased, do not posses a variability of stimuli physical or semantic properties, or they only provide information about saliency and keep private or discard information about temporal order of fixations (scanpath). For this reasons, we propose to the scientific community an open project in which public data coming from different experiments of eye-tracking can be collected into a unique and easily transferable format, together with an open source software package for data usage, statistics calculation, and implementation of the most common metrics for saliency and scanpath prediction. The project is advised by the MIT Saliency Team [1] and re-

[1] http://saliency.mit.edu/datasets.html

sources are publicly available [2].

To deal with the fact that the images in the available datasets often have a strong semantic content, and in collaboration with the Policlinico alle Scotte di Siena we have collected SIENA12, a dataset of fixations of human subjects on 12 grayscale images. The images were selected to minimize the semantic content and for this reason they include natural landscapes and geometric content, both in abstract images and natural scenes.

B.2 SIENA12

The SIENA12 dataset includes 12 grayscale images. It was collected by the author of this paper in collaboration with EVALAB at Policlinico Alle Scotte in Siena. The images have been selected so as to minimize the semantic content of the scenes. Images include natural scenes, human constructions, but also abstract contents.

Dataset Name	SIENA12
Number of images	12
Size	1024x768 px
Categories	Outdoor, natural, synthetic
Number of observers	23
Age of the observers	From 23 to 52
Task	Free-viewing
Duration	5 seconds
Eye-tracker	ASL 504 (240 Hz)
Screen	LCD 1024 $\times$ 768 px (31 $\times$ 51 cm)
Eye-screen distance	72 cm
Other information	Grayscale images

Table B.1: **Tech. spec. of the dataset SIENA12**

[2] http://sailab.diism.unisi.it/fixatons/

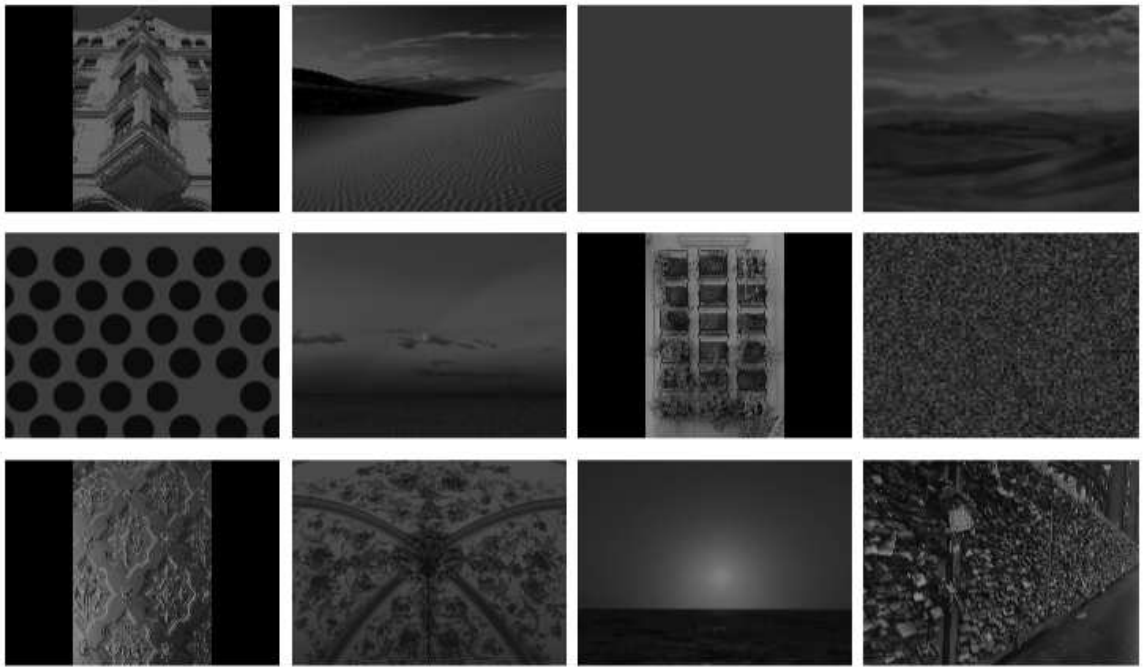


Figure B.1: **Images from SIENA12.** Images of Siena 12 have been properly selected to reduce semantic content as more as possible. The authors thank Danilo Pileri for kindly providing images of the dataset.

Protocol for data collection in SIENA12

Visual data are collected through a 240 Hz eye-tracking system (ASL 504, Applied Science Laboratories, Bedford, MA, USA) allowing for a remote tracking of the point of gaze on a calibrated surface (LCD screen, 1024 $\times$ 768 px, 31 $\times$ 51 cm). A chin-rest is used to maintain constant the relative distances between the eyes of the subject, the eye-tracker optics and the screen (eye/eye-trackers distance, 68 cm; eye/screen distance, 72 cm). The height of the chin-rest is set in order to get the line of sight of the subjects at the rest position perpendicular to the centre of the screen. The stimuli presentation and the data collection is managed by customized software. All recordings are conducted in complete darkness, measuring one eye.

A nine-point calibration is performed trough an interactive user interface provided by the manufacturer. The operator instructs the participant to perform 5 seconds tasks of free visual exploration. A total of 12 images is presented. Between one image and another, a central dot (of about 2 seconds) is displayed. The order of the images composing the sequence of the experimental stimuli is randomly chosen to prevent bias.

Raw data has been processed with the publicly available python library PyGaze-Analyser [19] in order to extract information about fixations.

B.3 Other datasets included in the collection

FixaTons include a collection of publicly available datasets as well as data collected in collaboration with Policlinico alle Scotte, University of Siena.

MIT1003

The dataset MIT1003 [42] is a large database of eye tracking data. The images, eye tracking data, and accompanying code in Matlab are all available on the web. This dataset can be used as training data for the MIT300 [41] benchmark since they share the same technical specification. MIT300 is not included in FixaTons collection because its data is kept private for a fair evaluation of the benchmark. More details about the protocol used for data collection are given in the referred paper.

Dataset Name	MIT1003
Number of images	1003
Size	Min. dim.: 405 px – Max. dim.: 1024 px
Categories	Landscape and portrait images
Number of observers	15
Age of the observers	From 18 to 35
Task	Free-viewing
Duration	3 seconds
Eye-tracker	ETL 400 ISCAN (240Hz)
Screen	LCD 1024 $\times$ 768 px (40.5 $\times$ 30 cm)
Eye-screen distance	75 cm
Other information	It can be used as training data for MIT300 benchmark

Table B.2: **Tech. spec. of the dataset MIT1003**

TORONTO

The dataset TORONTO [11] has been presented together with the AIM model (Attention based on Information Maximization). Observers age is not specified but they have been selected between undergraduate and graduate students. A large portion of images here do not contain particular regions of interest. More details about the protocol used for data collection are given in the referred paper.

Dataset Name	TORONTO
Number of images	120
Size	Min. dim.: 681x511 px
Categories	Outdoor and indoor scenes
Number of observers	20
Task	Free-viewing
Duration	4 seconds
Eye-tracker	ERICA workstation
Other information	It can be used as training data for MIT300 benchmark

Table B.3: **Tech. spec. of the dataset TORONTO**

KOOTSTRA

The KOOTSTRA [46] dataset is a collection of complex content photographic images. A total of 99 photographic images in five different categories were presented to the participants. Nineteen of them have been selected explicitly for containing

symmetrical natural objects (flowers or plants). The five categories span over a wide variety of cultural, natural, geometrical content.

Dataset Name	KOOTSTRA
Number of images	99
Size	Min. dim.: 1024x768 px
Categories	flowers, animals, street scenes, buildings, outdoor natural
Number of observers	31
Age of the observers	From 17 to 32
Task	Free-viewing
Duration	5 seconds
Eye-tracker	Eyelink I head-mounted eye-tracking system (SR research)
Screen	LCD 1024x768 px (36x27 cm)
Eye-screen distance	70 cm

Table B.4: **Tech. spec. of the dataset KOOTSTRA**

B.4 Online resources

- Webpage of the project: `http://sailab.diism.unisi.it/fixatons/`
- Data download: `https://drive.google.com/open?id=1TQSaq5J0p_oCdkyVZ-IzBltLwJ2cm3UA`
- Software library: `https://github.com/dariozanca/FixaTons`

B.5 Structure of the FixaTons collection

- FixaTons
  - DATASET_NAME
    - STIMULI : contains original images. They can have different file format (jpg, jpeg, png,...)
    - SCANPATHS : contains one folder for each image
      - IMAGE_ID : it contains one file for each scanpath of that image scanpaths are matrices rows of this matrices describe fixations each fixation is of the form : [x-pixel, y-pixel, initial time, final time]. Times are in seconds.
    - FIXATION_MAPS : contains a fixation map of each original image they are matrices of zeros (non-fixated pixels) and ones (fixated pixels). They can have different file format (jpg, jpeg, png,...)

* SALIENCY_MAPS : contains saliency maps of each original image they are generated from human data. They can have different file format (jpg, jpeg, png,...)

B.6 Software included

Some software tools are provided together with the collection for an easy use and visualization of the data.

Software is written in python. All the functions are included in the file *FixaTons.py*. They make use of the public library OpenCV which should be installed on the machine before the use of *FixaTons.py*.

Functions can be divided in five main categories:

- List information
- Get data (matrices)
- Visualize data
- Compute metrics.
- Compute statistics

List information

The collection comprehend different datasets, each of them with different stimuli names, number of subjects, subjects id's, etc. The provided software allows to easily get this kind of information.

- **FixaTons.list.dataset**(): This functions returns a list with the names of the datasets included in the collection.
- **FixaTons.list.stimuli**(**DATASET_NAME**): This functions lists the names of the stimuli of a specified dataset.
- **FixaTons.list.subjects**(**DATASET_NAME, STIMULUS_NAME**): This functions lists the id's of the subjects which have been watching a specified stimuli of a dataset.

Get data (matrices)

Different functions allows to get data in form of numpy array.

- **FixaTons.get.stimulus(DATASET_NAME, STIMULUS_NAME)**: This functions returns the matrix of pixels of a specified stimulus. Notice that, both DATASET_NAME and STIMULUS_NAME need to be specified. The latter, must include file extension. The returned matrix could be 2- or 3-dimensional.
- **FixaTons.get.fixation_map(DATASET_NAME, STIMULUS_NAME)**: This functions returns the matrix of pixels of the fixation map of a specified stimulus. Notice that, both DATASET_NAME and STIMULUS_NAME need to be specified. The latter, must include file extension. The returned matrix is a 2-dimensional matrix with 1 on fixated locations and 0 elsewhere.
- **FixaTons.get.saliency_map(DATASET_NAME, STIMULUS_NAME)**: This functions returns the matrix of pixels of the saliency map of a specified stimulus. Saliency map has been obtained by convolving the fixation map with a proper gaussian filter (corresponding to one degree of visual angle). Notice that, both DATASET_NAME and STIMULUS_NAME need to be specified. The latter, must include file extension. The returned matrix is a 2-dimensional matrix.
- **FixaTons.get.scanpath(DATASET_NAME, STIMULUS_NAME, subject = 0)**: This functions returns the matrix of fixations of a specified stimulus. The scanpath matrix contains a row for each fixation. Each row is of the type [*x, y, initial_t, final_time*]. By default, one random scanpath is chosen between available subjects. For a specific subject, it is possible to specify its id on the additional argument subject=id.

Visualize data

For an easy visualization of the data, some functions have been included in the library.

- **FixaTons.show.map(DATASET_NAME, STIMULUS_NAME, showSalMap = True, showFixMap = False, plotMaxDim = 0)**: This functions uses cv2 standard library to visualize a specified stimulus. By default, stimulus is shown with its saliency map aside. It is possible to deactivate such option by setting the additional argument showSalMap=False. It is possible to show also (or alternatively) the fixation map by setting the additional argument showFixMap=True. Depending on the monitor or the image dimensions, it could be convenient to resize the images before to plot them. In such a case, user could indicate in the additional argument plotMaxDim=500 to set, for example, the maximum dimension to 500. By default, images are not resized.
- **FixaTons.show.scanpath(DATASET_NAME, STIMULUS_NAME, subject = 0, animation = False, putNumbers = True, plotMaxDim = 0)**: This functions

uses cv2 standard library to visualize the scanpath of a specified stimulus. By default, one random scanpath is chosen between available subjects. For a specific subject, it is possible to specify its id on the additional argument subject=id. It is possible to visualize it as an animation by setting the additional argument animation=True. Depending on the monitor or the image dimensions, it could be convenient to resize the images before to plot them. In such a case, user could indicate in the additional argument plotMaxDim=500 to set, for example, the maximum dimension to 500. By default, images are not resized.

Compute metrics

An implementation of the most common metrics to compute saliency maps similarity and scanpaths similarity is included in the software provided with FixaTons. A mathematical description of the metrics of scanpath similarity is postponed to the Appendix.

- Saliency Map similarities
  - **FixaTons.metrics.KLdiv(saliencyMap1, saliencyMap2)**: This function computes the Kullback–Leibler divergence between two continuous saliency maps.
  - **FixaTons.metrics.AUC_Judd(saliencyMap, fixationMap, jitter = True, toPlot = False)** Given a continuous saliency map (normally the output of a saliency model) and a fixation map (matrix with 1's at fixated locations, 0's elsewhere), it computes the Area Under the ROC curve, in the implementation described by Judd in [42].
  - **FixaTons.metrics.NSS(saliencyMap, fixationMap)** Given a continuous saliency map (normally the output of a saliency model) and a fixation map (matrix with 1's at fixated locations, 0's elsewhere), it computes the Normalized Scanpath Saliency
- Scanpaths similarities
  - **FixaTons.metrics.euclidean_distance(human_scanpath, simulated_scanpath)**: This function computes the euclidean distance between two scanpaths. More details are given on the Appendix.
  - **FixaTons.metrics.string_edit_distance(stimulus, human_scanpath, simulated_scanpath, n = 5, substitution_cost=1)**: This function computes the string-edit distance between two scanpaths. More details are given on the Appendix.

- **FixaTons.metrics.time_delay_embedding_distance(human_scanpath, simulated_scanpath, k = 3, distance_mode = 'Mean')**: This function computes the time-delay embedding distance between two scanpaths. More details are given on the Appendix.
- **FixaTons.metrics.scaled_time_delay_embedding_distance(human_scanpath, simulated_scanpath, image, toPlot = False))**: This function computes the scaled time-delay embedding distance between two scanpaths. More details are given on the Appendix.

Compute statistics

It is possible to compute statistics for the overall collection, or for a specific dataset, about some scanpath properties.

- **FixaTons.stats.statistics(DATASET_NAME=None)**: This functions returns a list with two values: fixations per second and the average of saccades length. If no dataset is specified, statistics are calculated on the whole FixaTons collection. To restrict computation on a specific dataset, it is sufficient to indicate its name on the additional argument DATASET_NAME.

Example of use

Here we propose a python script which show a complete example of use of some facilities.

```
# import the library
import FixaTons

#shuffle(dataset_list)
for dataset in FixaTons.list.datasets():

# For all images in that dataset
for image in FixaTons.list.stimuli(dataset):

# Show the image aside its saliency map (5 seconds dy default).
FixaTons.show.map(dataset, image, plotMaxDim=1500)

# Then, for all the subjects that watched that image,
for subject in FixaTons.list.subjects(dataset, image):

# Show the correspondent scanpath as an animation.
# (Look, time of exploration in the animation is the
# exact time, from the dataset.)
FixaTons.show.scanpath(dataset, image, subject,
animation=True,
```

```
21 plotMaxDim=1000,
22 wait_time=1000)
```

Listing B.1: Example of use. Complete Python script.

Appendix C

Publications

Journal papers

1. **Dario Zanca**, M. Gori, A. Rufa, "A Unified Computational Framework for Visual Attention Dynamics", *Progress in Brain Research*, vol. 248, 2018. **Candidate's contributions**: designed algorithms, carried out theoretical analyses, experimental setup.

Peer reviewed conference papers

1. **Dario Zanca**, M. Gori, "Variational Laws of Visual Attention for Dynamic Scenes", *Advances in Neural Information Processing Systems* (*NIPS 2017*), pages:3823–3832, 2017. **Candidate's contributions**: designed algorithms, carried out theoretical analyses, experimental setup.

Papers under review

1. **Dario Zanca**, S. Melacci, M. Gori, "Gravitational Laws of Focus of Attention", *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*TPAMI*). **Candidate's contributions**: designed algorithms, carried out theoretical analyses, experimental setup.

2. G. Marra, **Dario Zanca**, A. Betti, M. Gori, "Learning Neuron Non-Linearities with Kernel-Based Deep Neural Networks", *International Conference on Learning Representations* (*ICLR 2019*). **Candidate's contributions**: designed and carried out experiments.

Other

1. **Dario Zanca**, V. Serchi, P.Piu, F. Rosini, A. Rufa, "FixaTons: A collection of Human Fixations Datasets and Metrics for Scanpath Similarity", *ArXiv preprint*,

arXiv:1802.02534, 2018. **Candidate's contributions**: data collection, designed algorithms, created a software library for saliency and scanpath metrics computation.

2. **Dario Zanca**, M. Gori, "Visual Attention driven by Convolutional Features", *ArXiv preprint*, arXiv:1807.10576, 2018. **Candidate's contributions**: designed algorithms, carried out theoretical analyses, experimental setup.

3. F. Giannini, V. Laveglia, A. Rossi, **Dario Zanca**, A. Zugarini, "Neural Networks for Beginners. A fast implementation in Matlab, Torch, TensorFlow", *ArXiv preprint*, arXiv:1703.05298, 2017. **Candidate's contributions**: designed and illustrated algorithms by examples.

4. **Dario Zanca**, M. McGill, "Coarse-to-Fine Q-Learning for Object Localisation on VHR Images", http://sailab.diism.unisi.it/coarse-to-fine-q-learning/, *Internal report*, 2018.

Bibliography

[1] Abeles, D., Amit, R., and Yuval-Greenberg, S. (2018). Oculomotor behavior during non-visual tasks: The role of visual saliency. *PloS one*, 13(6):e0198242.

[2] Achanta, R., Estrada, F., Wils, P., and Süsstrunk, S. (2008). Salient region detection and segmentation. In *International conference on computer vision systems*, pages 66–75. Springer.

[3] Bergen, J. R., Anandan, P., Hanna, K. J., and Hingorani, R. (1992). Hierarchical model-based motion estimation. In *European conference on computer vision*, pages 237–252. Springer.

[4] Betti, A. and Gori, M. (2018). Convolutional networks in visual environments. *ArXiv preprint, arXiv:1801.07110*.

[5] Betz, T., Kietzmann, T. C., Wilming, N., and König, P. (2010). Investigating task-dependent top-down effects on overt visual attention. *Journal of vision*, 10(3):15–15.

[6] Borji, A. (2018). Saliency prediction in the deep learning era: An empirical investigation. *arXiv preprint arXiv:1810.03716*.

[7] 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.

[8] Borji, A. and Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*.

[9] Borji, A., Tavakoli, H. R., and Bylinskii, Z. (2018). Bottom-up attention, models of. *arXiv preprint arXiv:1810.05680*.

[10] 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.

[11] Bruce, N. and Tsotsos, J. (2007). Attention based on information maximization. *Journal of Vision*, 7(9):950–950.

[12] Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., and Torralba, A. (2015). Mit saliency benchmark.

[13] Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., and Durand, F. (2018). What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*.

[14] Cerf, M., Frady, E. P., and Koch, C. (2008). Using semantic content as cues for better scanpath prediction. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, pages 143–146. ACM.

[15] Choi, Y. S., Mosley, A. D., and Stark, L. W. (1995). String editing analysis of human visual search. *Optometry and vision science: official publication of the American Academy of Optometry*, 72(7):439–451.

[16] Connor, C. E., Egeth, H. E., and Yantis, S. (2004). Visual attention: bottom-up versus top-down. *Current biology*, 14(19):R850–R852.

[17] Cornia, M., Baraldi, L., Serra, G., and Cucchiara, R. (2016). A deep multi-level network for saliency prediction. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3488–3493. IEEE.

[18] Coutrot, A. and Guyader, N. (2013). Toward the introduction of auditory information in dynamic visual attention models. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*, pages 1–4. IEEE.

[19] Dalmaijer, E. S., Mathôt, S., and Van der Stigchel, S. (2014). Pygaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior research methods*, 46(4):913–921.

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

[21] Dupont, P., Orban, G., De Bruyn, B., Verbruggen, A., and Mortelmans, L. (1994). Many areas in the human brain respond to visual motion. *Journal of neurophysiology*, 72(3):1420–1424.

[22] Einhäuser, W., Spain, M., and Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18–18.

[23] Elgammal, A., Harwood, D., and Davis, L. (2000). Non-parametric model for background subtraction. In *European conference on computer vision*, pages 751–767. Springer.

[24] Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer.

[25] Feynman, R. P., Leighton, R. B., and Sands, M. (1965). The feynman lectures on physics; vol. i. *American Journal of Physics*, 33(9):750–752.

[26] 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):6–6.

[27] Frintrop, S. (2006). Vocus: a visual attention system for object detection and goal-directed search [ph. d. dissertation]. *Rheinische Friedrich-Wilhelms-Universitat, Bonn, Germany*.

[28] Garcia-Diaz, A., Leboran, V., Fdez-Vidal, X. R., and Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of vision*, 12(6):17–17.

[29] Gelfand, I. M., Silverman, R. A., et al. (2000). *Calculus of variations*. Courier Corporation.

[30] Gori, M., Maggini, M., and Rossi, A. (2016). Neural network training as a dissipative process. *Neural Networks*, 81:72–80.

[31] Hadizadeh, H., Enriquez, M. J., and Bajic, I. V. (2012). Eye-tracking database for a set of standard video sequences. *IEEE Transactions on Image Processing*, 21(2):898–903.

[32] Hainline, L., Turkel, J., Abramov, I., Lemerise, E., and Harris, C. M. (1984). Characteristics of saccades in human infants. *Vision research*, 24(12):1771–1780.

[33] Harel, J., Koch, C., and Perona, P. (2006). A saliency implementation in matlab. *URL: http://www. klab. caltech. edu/harel/share/gbvs. php*.

[34] Hirschmuller, H., Innocent, P. R., and Garibaldi, J. M. (2002). Fast, unconstrained camera motion estimation from stereo without tracking and robust statistics. In *Control, Automation, Robotics and Vision, 2002. ICARCV 2002. 7th International Conference on*, volume 2, pages 1099–1104. IEEE.

[35] Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, 17(1-3):185–203.

[36] Itti, L. and Baldi, P. F. (2006). Bayesian surprise attracts human attention. In *Advances in neural information processing systems*, pages 547–554.

[37] Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194.

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

[39] Jiang, M., Huang, S., Duan, J., and Zhao, Q. (2015). Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[40] Jones, E., Oliphant, T., and Peterson, P. (2014). {SciPy}: open source scientific tools for {Python}. *Online resources*.

[41] Judd, T., Durand, F., and Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. *MIT Reports*.

[42] Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE.

[43] Jurafsky, D. and Martin, J. H. (2014). *Speech and language processing*, volume 3. Pearson London.

[44] Koch, C. and Ullman, S. (1987). Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer.

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

[46] Kootstra, G., de Boer, B., and Schomaker, L. R. (2011). Predicting eye fixations on complex visual stimuli using local symmetry. *Cognitive computation*, 3(1):223–240.

[47] Kowler, E. (2009). Attention and eye movements. In Squire, L. R., editor, *Encyclopedia of Neuroscience*, pages 605 – 616. Academic Press, Oxford.

[48] Krauzlis, R. J. (2013). Chapter 32 - eye movements. In Squire, L. R., Berg, D., Bloom, F. E., du Lac, S., Ghosh, A., and Spitzer, N. C., editors, *Fundamental Neuroscience (Fourth Edition)*, pages 697 – 714. Academic Press, San Diego, fourth edition edition.

[49] Kruthiventi, S. S., Ayush, K., and Babu, R. V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456.

[50] 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*.

[51] Land, M. F. (1999). The human eye: Structure and function.

[52] Larson, A. M. and Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, 9(10):6–6.

[53] Le Meur, O. and Liu, Z. (2015). Saccadic model of eye movements for free-viewing condition. *Vision research*, 116:152–164.

[54] Lee, T. S. and Stella, X. Y. (2000). An information-theoretic framework for understanding saccadic eye movements. In *Advances in neural information processing systems*, pages 834–840.

[55] Maggini, M. and Rossi, A. (2016). On-line learning on temporal manifolds. In *Conference of the Italian Association for Artificial Intelligence*, pages 321–333. Springer.

[56] McMains, S. and Kastner, S. (2011). Interactions of top-down and bottom-up mechanisms in human visual cortex. *Journal of Neuroscience*, 31(2):587–597.

[57] Pashler, H. (2016). *Attention*. Psychology Press.

[58] Petzold, L. (1983). Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM journal on scientific and statistical computing*, 4(1):136–148.

[59] Pieters, R. and Wedel, M. (2007). Goal control of attention to advertising: The yarbus implication. *Journal of consumer research*, 34(2):224–233.

[60] Posner, M. I., Rafal, R. D., Choate, L. S., and Vaughan, J. (1985). Inhibition of return: Neural basis and function. *Cognitive neuropsychology*, 2(3):211–228.

[61] Privitera, C. M. and Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on pattern analysis and machine intelligence*, 22(9):970–982.

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

[63] Renninger, L. W., Coughlan, J. M., Verghese, P., and Malik, J. (2005). An information maximization model of eye movements. In *Advances in neural information processing systems*, pages 1121–1128.

[64] Rossi, A., Rizzo, A., and Montefoschi, F. (2018). Atm protection using embedded deep learning solutions. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 371–382. Springer.

[65] Schlingensiepen, K.-H., Campbell, F., Legge, G., and Walker, T. (1986). The importance of eye movements in the analysis of simple patterns. *Vision Research*, 26(7):1111–1117.

[66] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626.

[67] Stoer, J. and Bulirsch, R. (2002). *Introduction to numerical analysis*. Texts in applied mathematics. Springer.

[68] Sunaert, S., Van Hecke, P., Marchal, G., and Orban, G. A. (1999). Motion-responsive regions of the human brain. *Experimental brain research*, 127(4):355–370.

[69] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

[70] 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.

[71] Tiezzi, M., Melacci, S., Maggini, M., and Frosini, A. (2018). Video surveillance of highway traffic events by deep learning architectures. In *International Conference on Artificial Neural Networks*, pages 584–593. Springer.

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

[73] Tynan, P. D. and Sekuler, R. (1982). Motion processing in peripheral vision: Reaction time and perceived velocity. *Vision research*, 22(1):61–68.

[74] 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.

[75] Wang, W., Chen, C., Wang, Y., Jiang, T., Fang, F., and Yao, Y. (2011). Simulating human saccadic scanpaths on natural images. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 441–448.

[76] Wright, R. D. and Ward, L. M. (2008). *Orienting of attention*. Oxford University Press.

[77] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

[78] Xu, M., Jiang, L., Sun, X., Ye, Z., and Wang, Z. (2017). Learning to detect video saliency with hevc features. *IEEE Transactions on Image Processing*, 26(1):369–385.

[79] Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer.

[80] Zanca, D. and Gori, M. (2017). Variational laws of visual attention for dynamic scenes. In *Advances in Neural Information Processing Systems*, pages 3823–3832.

[81] Zanca, D., Serchi, V., Piu, P., Rosini, F., and Rufa, A. (2018). Fixatons: A collection of human fixations datasets and metrics for scanpath similarity. *arXiv preprint arXiv:1802.02534*.

[82] Zhang, J. and Sclaroff, S. (2013). Saliency detection: A boolean map approach. In *Proceedings of the IEEE international conference on computer vision*, pages 153–160.

[83] 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.

[84] Zivkovic, Z. (2004). Improved adaptive gaussian mixture model for background subtraction. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 28–31. IEEE.