Modelling task-dependent eye guidance to objects in pictures

Cognitive Computation manuscript No.
(will be inserted by the editor)

Modelling task-dependent eye guidance to objects in pictures

Antonio Clavelli · Dimosthenis Karatzas · Josep Lladós · Mario Ferraro · Giuseppe Boccignone

Received: date / Accepted: date

Abstract We introduce a model of attentional eye guidance based on the rationale that the deployment of gaze is to be considered in the context of a general action-perception loop relying on two strictly intertwined processes: sensory processing, depending on current gaze position, identifies sources of information that are most valuable under the given task; motor processing links such information with the oculomotor act by sampling the next gaze position and thus performing the gaze shift. In such a framework, the choice of where to look next is task-dependent and oriented to classes of objects embedded within pictures of complex scenes. The dependence on task is taken into account by exploiting the value and the payoff of gazing at certain image patches or proto-objects that provide a sparse representation of the scene objects. The different levels of the action-perception loop are represented in probabilistic form and eventually give rise to a stochastic process that generates the gaze sequence. This way the model also accounts for statistical properties of gaze shifts such as individual scan path variability. Results of the simulations are compared either with experimental data derived from publicly available datasets and from our own experiments.

Keywords Visual attention · Gaze guidance · Value · Payoff · Stochastic fixation prediction

A. Clavelli · D. Karatzas · J. Lladós
Computer Vision Center, Universitat Autonoma de Barcelona
Edifici O, Campus UAB, 08193 Bellaterra (Cerdanyola) Barcelona, Spain
E-mail: {aclavelli,dimos,josep}@cvc.uab.cat

M. Ferraro
Dipartimento di Fisica, Universitá di Torino
via Pietro Giuria 1, 10125 Torino, Italy
E-mail: ferraro@ph.unito.it

G. Boccignone
Dipartimento di Informatica, Universitá di Milano
via Comelico 39/41, 10135 Milano, Italy
E-mail: giuseppe.boccignone@unimi.it

1 Introduction

This paper presents a probabilistic computational model of eye guidance for task-dependent attention deployment to objects in semantically rich pictures of natural scenes.

In the field of psychology, there exists a wide variety of theories and models on visual attention (see, e.g., the review by Heinke and Humphreys [46]). Among the most influential for computational attention systems, the well known Treisman's Feature Integration Theory (FIT) [106,105], Wolfe's Guided Search Model [115] aiming at explaining and predicting the results of visual search experiments, Desimone and Duncan's Biased Competition Model (BCM, [28]), Rensink's triadic architecture [83], and the Koch and Ullman's bottom-up model [59].

Other psychophysical models have addressed attention modelling in a more formal framework. One notable example is Bundensen's Theory of Visual Attention (TVA, [17]), further developed by Logan into the CODE theory of visual attention (CTVA, [64]). Also, theoretical approaches to visual search have been devised by exploiting Signal Detection Theory [76].

At a different level of explanation, other proposals have been conceived in terms of connectionist models, such as MORSEL (Multiple Object Recognition and attentional SELection, [69]), SLAM (SeLective Attention Model) [80], SERR (SEarch via Recursive Rejection) [52], and SAIM (Selective Attention for Identification

Model by Heinke and Humphreys [45]) subsequently refined in the Visual Search SAIM (VS-SAIM) [44].

To a large extent, the psychological literature was conceived and fed on simple stimuli, nevertheless the key role that the above models continue to play in understanding attentive behaviour should not be overlooked. For example, many current computational approaches, by and large, build upon the bottom-up salience based model by Itti *et al.* [54], which in turn is the computational counterpart of Koch and Ullman and Treisman's FIT models. The seminal work of Torralba et al. [104], draws on an important component of Rensink's triadic architecture [83], in that it considers contextual information such as gist - the abstract meaning of a scene, e.g., a city scene, etc. - and layout - the spatial arrangement of the objects in a scene. More recently, Wischnewski *et al.* [114] have presented a computational model that integrates Bundensen's TVA [17].

However, in the last three decades, psychological models have been adapted and extended in many respects, within the computational vision field where the goal is to deal with attention models and systems that are able to cope with natural complex scenes rather than simple stimuli and synthetical images (e.g., see [38] and the most recent review by Borji and Itti [14]). The adoption of complex stimuli has sustained a new brand of computational theories, though this theoretical development is still at an early stage: up to this date, nobody has really succeeded in predicting the sequence of fixations of a human observer looking at an arbitrary scene [38]. This is not surprising given the complexity of the problem. One might think that issues of generalisation from simple to complex contexts are nothing more than a minor theoretical inconvenience; but, indeed, the generalisation from simple to complex patterns might not be straightforward. As it has been noted in the case of attentive search, a model that exploits handpicked features may fail utterly when dealing with realistic objects or scenes [117]. More precisely, the aim of a computational model of visual attention is to answer the question *Where to Look Next?* by providing: 1) at the *computational theory level* (in the sense of Marr, [66]; defining the input/output computation at time t), an account of the mapping from a complex natural scene, say $\mathbf{I}$, to a sequence of gaze locations $\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots$, under a given task $\mathbf{T}$, namely

$$\mathbf{I} \underset{\mathbf{T}}{\longmapsto} \{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}, \tag{1}$$

where the sequence $\{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}$ can be used to define a scan path; 2) at the *algorithmic level*, [66], a procedure that simulates such mapping (we will not specifically address here the third level of neural realisation [66]).

Current approaches within this field suffer from a number of limitations: they mostly rely on a low-level salience based representation of the visual input, they seldom take into account the task's role, and eventually they overlook the eye guidance problem, in particular the actual generation of gaze-shifts (cfr. Sect. 2 for a wider discussion, but see Tatler *et al* [100] for a lucid critical review of current methods).

Thus, the goal and the novelty of the study presented here is to propose an integrated computational model that: i) accounts for task dependent attentive processing of complex natural images by exploiting multiple levels of representation of the visual input; ii) describes statistical properties of gaze shifts performed by the "foraging eye" as closely as possible, including inter-individual scan path variability.

The rationale behind our approach is that the deployment of gaze to regions of an image is to be considered in the context of a general action-perception loop [39,89] relying on two strictly intertwined processes: sensory processing, depending on current gaze position $\mathbf{r}_F(t)$, identifies sources of information that are most valuable under the given task; motor processing links such information with the oculomotor act by sampling the next gaze position $\mathbf{r}_F(t+1)$ and thus performing the gaze shift $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t+1)$ [40]. The new gaze position $\mathbf{r}_F(t+1)$ provides a novel sight to sense the scene and the loop starts over, until the task is fulfilled.

We embrace the view that visual attention (and in particular overt attention) derives from the activity of such a sensorimotor loop, and implements a specialized form of *decision* based on the utility or value of information as framed by the given task. It is important to make clear from the beginning that our use of the term decision accounts for a decision function or estimator evaluated under a utility function (or, equivalently, a loss function) as technically understood in statistical decision theory and in particular in Bayesian decision theory [85]. Thus, in this respect such term should not be generally intended as cognitive or conscious decision. In a foraging metaphor (see [12] for an in-depth technical presentation and a recent paper by Wolfe [116] relating foraging to visual search), the eye is a forager that feeds on valuable information. The forager, moment to moment, is confronted with the choice between "feed" - that is, performing local intensive exploration (fixational eye movements) of the current patch of the attentional landscape -, or "fly", by making an extensive relocation (saccade) toward a new patch. This choice, in turn, entails the decision on where to go next based on maximising the expected payoff (or minimising the average risk) under the given task.

Here, we model how these processes and the different levels of representation (proto-objects, objects, value, [89]) may interact to fulfil task demands. In order to account for the several latent factors involved in the guidance of eye-movements [89] - for example, oculomotor biases and the "internal" noise [102] -, we assume that such representations are shaped in the form of probability distributions and that the moment-to-moment relocation of gaze, the walk of the foraging eye on the attentional landscape, is generated by an underlying stochastic process.

We will refer to two tasks, free-viewing and search for some kind of object, which will tune the action-perception loop generating the gaze shift. Also, we will consider, within the uncountable variety of classes of objects that can occur in real-world scenes, text and faces since these are known to attract attention even along a free-viewing task [20,112].

The paper is organised as follows. Sect. 2 provides background and rationales for setting up the model. The latter is introduced in Sect. 3 where we present the working assumptions and detail the action-perception loop, in terms of a probabilistic framework accounting for the different representational levels (proto-objects, objects, value, task) involved in the perceptual and oculomotor processes.

The model proposed here is then simulated (Sect. 4) by resorting to a stochastic sampling procedure derived from the Ecological Sampling (ES) approach [12].

In Sect. 5, simulation results are compared either with experimental data derived from publicly available datasets and with data eye-tracked from human subjects in our own experiments. An overall discussion is finally presented in Sect. 6.

2 Background and rationales

The primary motivation to engage in such a challenging program is that most current approaches in computational modelling of attention share a number of limitations. To summarize the discussion provided in this Section, we make explicit the common practice of computational approaches to conceive the mapping (1), as a two step procedure: first, obtain a suitable representation $\mathcal{R}$, i.e., $\mathbf{I} \underset{\mathbf{T}}{\mapsto} \mathcal{R}$; second, use $\mathcal{R}$ to generate the scan path, $\mathcal{R} \underset{\mathbf{T}}{\mapsto} \{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}$.

2.1 Levels of representation and control

For what concerns the first step, the guidance of eye movements is likely to be influenced by a hierarchy of several interacting control loops, operating at different levels of processing of the whole action-perception loop. Each processing level exploits the most suitable representation of the viewed scene for its own level of abstraction: Schütz, *et al.* [89], in a plausible portrayal, have singled out salience, objects, values, and plans. Up to this date, the majority of computational models have retained a central place for low-level visual conspicuity or early salience [100,14].

The representation $\mathcal{R}$ is typically shaped in the form of a spatial saliency map, which is mostly derived bottom-up, following Itti *et al* [54] (e.g., see the most recent review [14]). The weakness of the bottom-up approach has been largely discussed (see, e.g. [100,37,32]). Indeed, the effect of early saliency on attention is likely to be a correlational effect rather than an actual causal one [37,89], though salience may be still more predictive than chance while preparing for a memory test as discussed by Foulsham and Underwood [37]. Two examples are provided in Fig. 1, where, as opposed to human scan paths (left images, free-viewing conditions), the scan paths generated by using a salience-based representation [54] (right images) only cover semantically important objects (text and faces) when these define - or are located nearby - regions of high contrast in colour, texture and luminance.

However, Torralba *et al.* [104] have shown that using prior knowledge on the typical spatial location of the search target, as well as contextual information (the gist of a scene, [83]) to modulate early saliency improves its fixation prediction. In a similar vein, object knowledge can be used to top-down tune early salience. In particular, when dealing with faces within the scene, a face detection step can provide a reliable cue to complement early conspicuity maps, as it has been shown by Cerf *et al* [21], deCroon *et al* [27], Marat *et al* [65], or a useful prior for Bayesian integration with low level cues [13]. This is indeed an important issue since faces may drive attention in a direct fashion [20].

More likely, early salience has only an indirect effect on attention, acting through recognised objects: observers attend to interesting objects and saliency contributes little extra information to fixation prediction[32]. Indeed, objects and their semantic value have been deemed as fundamental for visual attention and eye guidance (e.g., [69,17,83,44], but see Scholl [88] for a review).

As discussed by Einhäuser *et al.* [32], objects predict fixations in individual images better than early salience. Surprisingly, they are rarely taken into account in computational models [100]. There are of course exceptions to this state of affairs, most notable ones those provided by Rao *et al.* [82], Sun *et al.* [98], the Bayesian models discussed by Chikkerur *et al.* [24] and Borji *et al.* [15].

In this paper we are not much involved in discussing neurobiological underpinnings of computational theories, but, interestingly enough the approach of fusing object-based information with low-level salience, either through straightforward combination [21,65] or in the formal framework of Bayesian modelling [13,24,15] provides a computational account of the way the lateral intraparietal area (LIP) of posterior parietal cortex acts as a *priority map* to guide the allocation of covert attention and eye movements (overt attention). The LIP is a cortical area located at the interface between visual input and oculomotor output and it is well known that LIP activity is biased by both bottom-up stimulus driven factors and top-down cognitive influences. Due to its role, LIP has been viewed as a sort of "final path" for saccade motor decisions. However LIP studies of visuo-spatial attention have shown that in addition to its saccade-related activity, its neurons exhibit robust responses to salient or task-relevant stimuli that are not targets of a gaze shift [40]. Thus, it has been proposed that that LIP encodes a stage of visual selection that communicates with but is distinct from a stage of motor selection [40]. The connection between the priority map, at the computational level and LIP, at the neural implementation level, has been explicitly put forward by Chikkerur *et al.* [24].

Clearly in order to posit objects as the unit of attention the concept of object must be known to the system. In vision science discussion has been devoted to entities that have come to be known as proto-objects or pre-attentive objects [83,88] since they need not to correspond exactly with conceptual or recognisable objects. These are intermediate entities between localised features and objects. Instead, they reflect the visual system's grouping of parts of the retinal input which are likely to correspond to parts of the same object in the real world. One suitable account for such issues is provided by the TVA-based model proposed by Wischnewski *et al.* [114].

As a matter of fact, in the real world, most fixations are on task-relevant objects and this may or may not correlate with the saliency of regions of the visual array [18,86]. For instance, the eye guidance process is quite different when an observer is engaged in a search task as opposed to a generic picture viewing task: it is well known that even though both bottom-up and top-down sources of information are available to search, the bottom-up information is largely ignored [117]. Further, when the behavioural task is manipulated, feature-based models can fail almost completely, as it has been shown by Einhäuser *et al.* [31], Foulsham and Underwood [37,107,108].

Different studies have recently taken into account the role of task assignment to observers. For purpose of object recognition or search, some authors use the given task (e.g., specified through key words [72]) for either directly biasing the saliency map toward known image features of the corresponding target object or for tuning the task-dependent attentional weights for proto-objects [114]. Top down weight tuning has a long tradition in the psychological literature of search where models such as FIT [105], Guided Search [115] and BCM (Desimone & Duncan, 1995) have largely concentrated on biasing the feature maps or proto-objects [17] in a global way to facilitate efficient search. However, again, in complex natural scenes the selection of features is far from trivial due to the high-dimensionality of the feature space and it is not unlikely that features be shared by the target and many distractors. Further, model's operations should be relatively stimulus independent: if two different classes of stimuli require different sets of parameters, and these must be supplied by the user in an unprincipled way, the model cannot be described as general. Alternatively, a measure of visual similarity or match between the gazed region and the search target has been proposed [117], although in this case one has to deal with the classic issues raised in the object recognition realm, for example, object pose variations and illumination changes.

However, this way of conceiving top-down influence turns to be a rather poor account when dealing with semantically rich natural images. For instance, even when a rather specific task is assigned - e.g., searching for objects of a specific class -, yet, objects of a different class may still act as distractors due to their intrinsic value or *motivational salience* [19] for the observer.

A more convenient way of accounting for this problem stems from the general rationale that the selection of stimuli by attention has important implications for the survival and wellbeing of an organism, and attentional priority reflects the overall value of such selection (see Anderson [1] for a recent discussion). In this perspective the assignment of a task to the observer implicitly defines a value for every point of the space, in the sense that information in some points is more relevant than in others for the completion of the task; the shifting of the gaze on a particular point, in turn, determines the payoff that can be gained. The payoff then is nothing else that the value, with respect to the completion of the task, obtained by moving the fovea in a given position. Thus points associated with high values produce, when fixated, high payoffs since these fixations bring the observer closer to her/his goal. This definition of payoff is similar to the broad definition of reward proposed by Maunsell [68]: "one such definition

would include not only the immediate primary rewards, but also other factors: the preference for a novel location or stimulus, the satisfaction of performing well or the desire to complete a given task." Such definition is consistent with the different psychological facets of reward: i) learning (including explicit and implicit knowledge produced by associative conditioning and cognitive processes); ii) affect or emotion (implicit liking and conscious pleasure); iii) motivation (implicit incentive salience wanting and cognitive incentive goals) [6].

To sum, such a payoff is an operational concept for describing the value that an observer, consciously or unconsciously, ascribes to an object, a behavioural act, or an internal physical state; the given task modulates the value assigned to a certain class (or classes) of objects that nevertheless compete with other objects of different motivational salience, so that the final oculomotor act is taken to maximise the expected gain.

There is a number of psychological and neurobiological studies showing the availability of value maps and loci of reward influencing the final gaze shift [81,63, 53,47]. Nevertheless, while salience, proto-objects and objects are representations that have been largely addressed in the context of human eye movements, albeit with different emphasis, in contrast value has been neglected until recently [89]. One reason is that in the real world there is seldom direct payoff (no orange juice for a primary reward) for making good eye movements or punishment for bad ones. The high attentional priority of ecologically pertinent stimuli can also be explained by mechanisms that do not implicate learning value through repeated pairings with reward. For example, a bias to attend to socially relevant stimuli is evident from infancy consistent with an inherited attentional bias that precedes learning [1].

Yet, developing eye guidance models based on reward is a difficult endeavour and computational models that use reward and uncertainty as central components are still in their infancy (but see the discussion by Tatler *et al.* [100]). Nevertheless, the effort shows his inner worth in that, by accounting for the many aspects of "biological value" - salience, significance, unpredictability, affective content - , it paves the way to a broader and more abstract dimension of information processing, as most recent results on the affective modulation of the visual processing stream advocate [77, 78], and to the effective exploitation of computational attention models in the emerging domain of social signal processing [109].

Fig. 1 Scan paths generated by free viewing pictures embedding semantic objects (faces, text). Left: scan paths obtained eye-tracking a human observer. Right: scan paths simulated using the Itti et *et. al* model [54] as implemented in the latest version of the saliency tool box downloaded from http://www.saliencytoolbox.net using the default parameters.

2.2 Generation of gaze shifts

The second step, that is $\mathcal{R} \mapsto \{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}$, brings in the question of *how* we look rather than *where*, an issue that is seldom taken into account in computational modelling. Actually, what can be shown, by analyzing the statistics of gaze shifts, is that there are systematic statistical tendencies in the oculomotor behaviour that are either common across observers [3,75,101,16] or specific for individuals [102] or specific classes of observers (see, e.g. the review by Toh *et. al* [103] on visual scanning strategies in psychotic, anxiety, and mood disorders and the remarkable recent study by Sprenger and colleagues concerning patients with schizophrenia [95]).

In most computational models $\mathcal{R}$ is usually evaluated in terms of its capacity for predicting the image regions that will be explored by covert and overt attentional shifts according to some evaluation measure [14]. In other cases, if needed for practical purposes, e.g. for robotic applications, the problem of oculomotor action selection is solved by adopting some deterministic choice procedure that usually relies on selecting the gaze position $\mathbf{r}$ as the argument that maximizes a measure on the given representation $\mathcal{R}$ (in brief, see [111] for using the $\arg\max_{\mathbf{r}} \mathcal{R}$ operation and [12,100], for an in-depth discussion), such as the maximum value of the saliency map [54] or the proto-object with the highest attentional weight [114]. This despite of the fact that Tatler and Vincent in their elegant study [102] found striking evidence that a model based on oculomotor tendencies alone performs better than the stan-

dard salience model. Further, they have shown that exploiting these oculomotor biases, the performance of a salience model can be improved from 56% to 80% by including the probability of gaze shift directions and amplitudes. Unfortunately, they did not provide neither a formal characterisation of the distributions at hand, nor a computational procedure to generate gaze shifts, since they directly exploited histograms of saccade directions and amplitudes gathered from the participants to the experiment.

An interesting question is how such tendencies arise. Clearly, they are not purely motoric but their origin is likely to account for a range of sources, from high-level knowledge or uncertainty about the structure of the visual environment and about the distribution of objects within the environment to low-level biomechanics [102]. It is worth noting that uncertainty comes into play since the earliest stage of visual processing: the human retina evolved such that high quality vision is restricted to the small part of the retina (about $2^0 - 5^0$ degrees of visual angle) aligned with the visual axis, the fovea at the centre of vision; for many visually-guided behaviours the coarse information from peripheral vision is insufficient (for a review see Strasburger *et al.* [97]). In certain circumstances, uncertainty may promote almost "blind" visual exploration strategies [102, 75], much like the behaviour of a foraging animal exploring the environment under incomplete information; indeed when animals have limited information about where targets (e.g., resource patches) are located, different random search strategies may provide different chances to find them [4]. On the other hand, motor biases in the oculomotor system are likely to promote small amplitude gaze shift rather than large amplitude saccades. Thus, amplitudes show a positively skewed, long-tailed distribution in most experimental settings in which complex scenes are viewed [99,101,102]. Failing to account properly for such characteristics results in scan path patterns that are fairly different from those generated by human observers (which can be easily noticed in the example provided in Fig. 1) and eventually in distributions of saccade amplitudes that do not match human eye behaviour.

More generally, randomness in motor responses originates from endogenous stochastic variations that affect each stage between a sensory event and the motor response sensing, information processing, movement planning and executing [5]. As Canosa put it, where we choose to look next at any given moment in time is not completely deterministic, but neither is it completely random [18], and the language of probability and stochastic processes [102,16,12] provides a principled framework to handle such an issue.

All together these factors nourish the variability that typically characterises scan paths produced by human observers. Indeed, when looking at static images or natural movies the moment-to-moment relocation of gaze is different among observers, even though the same locations are taken into account, a long standing issue that recently has been soundly investigated by Dorr and colleagues [30] in their experimental work. Notably, the variations in individual scan paths still hold when the scene contains semantically rich "objects".

Many works have addressed the problem of measuring the similarity of scan paths produced by different subjects - or the same subject in different trials - observing the same scene under the same task (a lively research line, see the discussion by Dewhurst *et al.* [29]). In contrast, the problem of modelling the variability of visual scan paths produced by human observers has hitherto been overlooked by most computational accounts [89,14]. Few works have been trying to cope with the variability issue, after the early work by Stark and colleagues [33,43]. Kimura *et al.* [56] have incorporated simple eye-movements patterns as a probabilistic prior; Ho Phuoc *et al.* [48] embed at least one parameter suitable to be tuned to obtain different saccade length distributions on static images, though statistics obtained by varying such parameter are still far from those of human data; others try to capture eye movements randomness [55,87] but limiting to specific tasks such as conjunctive visual search. A few more exceptions can be found, but only in the very peculiar field of eye-movements in reading (see Feng for a discussion[35]). More recently, the variability issue has been explicitly addressed in the theoretical context of Lévy flights [16, 10,96] and composite α-stable random walks [11,12], however the perceptual component was limited to a minimal core (e.g., based on a bottom-up salience map) sufficient enough to support the eye guidance component.

In the model presented in the following Section, we attempt at filling such gaps at both the representational and the scanning strategy levels.

3 The model

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

$\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t+1)$ is a transition whose dynamics is that of a stochastic process.

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

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

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

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

The conditional dependencies between RVs $\mathcal{A}(t), \mathcal{A}(t+1), \mathcal{W}(t), \mathcal{W}(t+1)\mathbf{r}_F(t+1), \mathbf{r}_F(t+1), \mathbf{T}, \mathbf{I}$ can be explicitly represented by means of the Probabilistic Graphical Model (PGM) depicted in Fig. 2,

A PGM [60] is a graph-based representation where nodes denote RVs and arrows code conditional dependencies between RVs. It is important to note that arrows do not generally represent causal relations, though in specific situations it could be the case. More precisely, the structural dependency $X \rightarrow Y$, states the probabilistic dependency of RV Y on X represented via the conditional probability $P(Y|X)$).

Indeed, this is one suitable way of formalising a model at the computational theory level [57].

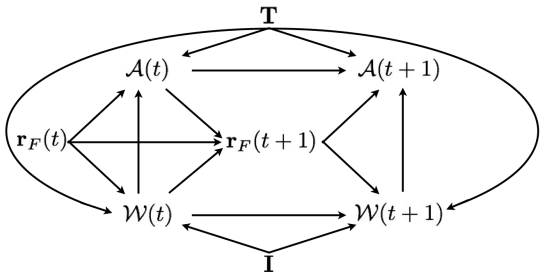


Fig. 2 The model represented as a dynamic Probabilistic Graphical Model. $\mathcal{A}(t)$ stands for the ensemble of time-varying random variables (RVs) defining the oculomotor action setting (for short, the *action* component); $\mathcal{W}(t)$ is the ensemble of time-varying RVs characterising the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

Note that the scheme in Fig. 2 can be read as a dynamic (time-varying) PGM [60]. Further, it is important to note that the state transition dynamics of the RVs from time t to time $t+1$ only depends on the state of such RVs a time t. In the language of stochastic processes this statement characterises a first order Markov process. Such formal assumption, which is largely exploited in dynamic PGMs [60] is occasionally summarised as a memoryless assumption about the process. By analogy with the psychological literature, this would amount to say that our model when used to perform a search task, implements a kind of visual search that has no memory [50]. However, such liberal interpretation turns to be improper. A first order Markov assumption about the gaze shift $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t+1)$ only states that the transition probability has the following property: $P(\mathbf{r}_F(t+1)|\mathbf{r}_F(t)) = P(\mathbf{r}_F(t+1)|\mathbf{r}_F(t), \mathbf{r}_F(t-1), \mathbf{r}_F(t-2), \cdots)$, namely, at time t the probability of the transition $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t+1)$ can be computed by conditioning on $\mathbf{r}_F(t)$, and earlier terms - at times $t-1, t-2, \cdots$ - need not be taken into account. The same holds for $P(\mathcal{W}(t+1)|\mathcal{W}(t), \mathbf{r}_F(t+1), \mathbf{I}, \mathbf{T})$ and $P(\mathcal{A}(t+1)|\mathcal{A}(t), \mathcal{W}(t+1), \mathbf{r}_F(t+1), \mathbf{T})$. However, as we will discuss later, there are RVs in the sets $\mathcal{W}(t), \mathcal{A}(t)$ that are used to define probability distributions over the image spatial support (for example, the priority map and the value map represented through the spatially

defined RVs $\mathbf{L}(t)$ and $\mathbf{V}(t)$, respectively) that, though evolving in time according to a first order Markov dynamics, keep track of events previously occurred. Thus, when engaged in a search task the gaze sampling mechanism may behave very differently from a sampling with no memory (i.e., with replacement [79]).

To keep things simple we will consider two tasks: a general "free-view" task ($\mathbf{T} = FV$) and a "look for x" ($\mathbf{T} = S$) or search task. Hence $\mathbf{T}$ is a binary RV. It can be seen from Fig. 3, that the task variable $\mathbf{T}$, at any time t, influences either the perceptual ensemble $\mathcal{W}(t)$ and the action ensemble $\mathcal{A}(t)$. In brief, this will be obtained by conditioning on task, at the perceptual level, the prior probability of gazing at certain objects within the scene, while at the action level, the task will modulate the probabilities of the value and the payoff related to a possible oculomotor act. In the following sections, we will provide concrete examples of the top-down role played by the task variable $\mathbf{T}$. Further, we instantiate and discuss the actual RVs characterising the general representational levels that we have summarised through the ensembles $\mathcal{W}(t)$ and $\mathcal{A}(t)$, together with their dependencies. As a result, the PGM presented in Fig. 2 will be eventually specified in a full probabilistic model that we introduce in Fig.3 below.

For explanatory convenience, we will start our discussion from the representations underpinning the perceived scene $\mathcal{W}(t)$, as available by "freezing" the loop at time t (Fig. 3) when gaze is deployed at $\mathbf{r}_F(t)$. Nevertheless, it is important to note that in this article we are not committing to any specific visual procedure, inasmuch as it serves the purpose of supporting the computational theory of the integrated loop.

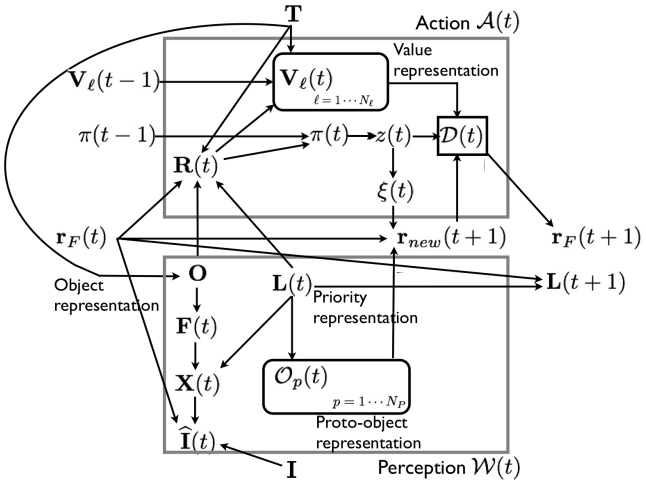


Fig. 3 A snapshot of the model when gaze is deployed at $\mathbf{r}_F(t)$. It provides a detailed view of the time slice t outlined in Fig. 2. Rounded boxes are "plates" denoting stacks of multiple random variables generated from the same distribution.

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

Consider the PGM specification of the model outlined in Fig. 3, and, in particular the perception component at the bottom of the scheme. At time t, the perceived scene $\mathcal{W}(t)$ is an ensemble of different representations, namely

- $\{\widehat{\mathbf{I}}(t), \mathbf{X}(t)\}$: the *visual front-end* given by the foveated image $\widehat{\mathbf{I}}$ and a local feature map $\mathbf{X}(t)$ [24];
- $\mathbf{L}(t)$: a *priority map*, that is a map of visual space constructed from a combination of properties of the external stimuli, intrinsic expectations, contextual knowledge [24,104];
- $\mathcal{O}(t)$: an ensemble of *proto-objects* or patches [12, 114,111];
- $\{\mathbf{O}, \mathbf{F}(t)\}$: an *object-level* representation, as determined by the classes of objects that can be embedded within the scene together with the visual features characterising the appearance of such objects [24]. In this study, we take into account faces and text regions that are known to attract attention even in a free viewing task [20,112], thus the RV accounting for objects is a binary one, *i.e.*, $\mathbf{O} = \{face, text\}$.

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

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

The "foraging eye", by gazing at $\mathbf{r}_F(t)$, allows the observer to gauge, at time t, the actual scene represented here by the given image $\mathbf{I}$ and thus to construct $\mathcal{W}(t)$. The first step for inferring the perceived scene $\mathcal{W}(t)$ is to derive a foveated representation of the input image $\mathbf{I}$. Many visual attention models do not take into account the retinal position of image information, and decreasing retinal acuity in the periphery is surprisingly overlooked [100]. Yet, retinal anisotropies in sampling play a role in tendencies to move the eyes in particular ways, and Tatler *et al.* [100] raised the point that the assumption of uniform spatial sampling can lead to distributions of saccade amplitudes that do not match human eye behaviour. Thus, in our model the starting point is represented by the foveated image $\widehat{\mathbf{I}}(t)$, that is $\mathbf{I}$ gazed at $\mathbf{r}_F(t)$. The foveated image $\widehat{\mathbf{I}}(t)$ is structured as a pair $\widehat{\mathbf{I}}(t) = \{\widehat{\mathbf{I}}_{LR}(t), \widehat{\mathbf{I}}_{HR}(t)\}$, respectively a low-resolution (LR) one, mainly exploited during long relocations of gaze, and a high resolution one (HR), mainly used to support local fixational movements and small saccades.

From the foveated image, perception is accomplished according to a hierarchical scheme (cfr., Fig. 3). The

structural dependencies shaping such hierarchy can be formalised in terms of probabilistic conditional dependencies among the RVs introduced above, which amounts to the following factorisation of the joint pdf introduced above:

$$
\begin{aligned}
&P(\mathbf{O},\mathbf{F}(t),\mathbf{L}(t),\mathbf{L}(t-1),\mathcal{O}(t),\\
&\qquad\qquad \mathbf{X}(t),\widehat{\mathbf{I}}(t)|\mathbf{I},\mathbf{T},\mathbf{r}_F(t),\mathbf{r}_F(t-1)) =\\
&P(\mathbf{O}|\mathbf{T})P(\mathbf{L}(t)|P(\mathbf{L}(t-1),\mathbf{r}_F(t-1))P(\mathcal{O}(t)|\mathbf{L}(t))\\
&\cdot P(\mathbf{F}(t)|\mathbf{O})\cdot P(\mathbf{X}(t)|\mathbf{L}(t),\mathbf{F}(t))\\
&\cdot P(\widehat{\mathbf{I}}(t)|\mathbf{r}_F(t),\mathbf{X}(t),\mathbf{I})
\end{aligned}
\tag{2}
$$

The factorization specified in Eq. 2 makes explicit the distributions at the different levels of representation from top to bottom: the object and object feature level, $P(\mathbf{O}|\mathbf{T})$ and $P(\mathbf{F}(t)|\mathbf{O})$, respectively; the priority map level, $P(\mathbf{L}(t)|P(\mathbf{L}(t-1),\mathbf{r}_F(t-1))$; the proto-object level, $P(\mathcal{O}(t)|\mathbf{L}(t))$; the local feature level that ties object features to prioritized locations, $P(\mathbf{X}(t)|\mathbf{L}(t),\mathbf{F}(t))$; the foveated image level $P(\widehat{\mathbf{I}}(t)|\mathbf{r}_F(t),\mathbf{X}(t),\mathbf{I})$.

Clearly, the probability of dealing with certain classes of objects, $P(\mathbf{O}|\mathbf{T})$ depends on the kind of images taken into account according to the task. The likelihood of spatially independent object-based features, i.e., $P(\mathbf{F}(t)|\mathbf{O})$, can be learned off-line with any suitable technique. Indeed, it is important to note that any perceptual inference model capable of top-down, object-based analysis and representation, can serve as a suitable one for the framework presented here, provided that a priority map $\mathbf{L}(t)$ is computed. One suitable procedure could be the one discussed by Chikkerur *et al.* [24], though in the work presented here there is a conceptual difference with respect to [24] in that we consider the generation of a sequence of gaze locations. Hence, the actual input to the visual inference process is in terms of a sequence of foveated images $\widehat{\mathbf{I}}(t)$. So, for instance the inference of the priority map becomes time and gaze dependent, i.e., $P(\mathbf{L}(t)|\widehat{\mathbf{I}}(t))$ rather than simply $P(\mathbf{L}|\mathbf{I})$.

The priority level representation can be inferred from the posterior $P(\mathbf{L}(t)|\widehat{\mathbf{I}}(t))$. Note that if the features $\mathbf{X}(t)$ are observed, then $\mathbf{L}(t)$ and $\mathbf{O}$ are conditionally dependent, and prioritization is biased by objects present in the scene. Note that, in the absence of object information, $P(\mathbf{F}(t)|\mathbf{O}) = P(\mathbf{F}(t))$ and $\mathbf{L}(t)$ boils down to a classic salience map. The attentional priority is related to both the object's salience and any top-down biases that influence the relative importance of that object to the subject, including the suppression of objects that have already been examined during visual search, thus playing a role in participating to the elusive Inhibition of Return (IOR) mechanism [113]. The reduction in the response to a stimulus that has been fixated essentially acts as a form of short term memory that lets the priority map keep track of which potential targets have been examined. This effect is here taken into account by letting the current $\mathbf{L}(t)$ to depend on gaze location and priority at time $t-1$, $P(\mathbf{L}(t)|\mathbf{L}(t-1),\mathbf{r}_F(t-1))$ (Fig. 3). This modelling choice is consistent with the finding that LIP neurons receive feedback about the selected action.

Note that the distribution on $\mathbf{L}$, $P(\mathbf{L}(t)|\mathbf{L}(t-1),\mathbf{r}_F(t-1))$, serves as a spatial prior to locate object features $\mathbf{F}$ on the early feature map $\mathbf{X}$. But, more generally, the priority map could also be used to take into account contextual spatial modulation of visual attention [104]. We do not consider here this problem, but integrating contextual issues in our scheme is readily done (say, in the form $P(\mathbf{L}(t)|\mathbf{L}(t-1),\mathbf{r}_F(t-1),Gist)$), and it has been experimented for a text localisation task in urban street pictures using an earlier and simplified version of the model presented here [26].

The time varying priority map $\mathbf{L}(t)$ is fundamental to organise a dynamic representation of the scene in terms of *proto-objects* [83,114,111,51], which serves as the actual dynamic support for gaze orienting. They are conceived as the dynamic interface between high-level and low-level processing, a "quick and dirty" interpretation of the scene [83] . There are several possibilities to compute a proto-object representation. One way is in compact form, from either a simple [111,51] or a more complex mid-level segmentation process (e.g., [114,8]); an alternative is to use a sparse representation [12]. This latter option, which we embrace, will be discussed in detail in Sec. 4.

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

Consider now the action component at the top of the PGM in Fig. 3. The oculomotor action setting $\mathcal{A}(t)$ under task $\mathbf{T}$ can be defined through the following ensemble of RVs:

- $\{\mathbf{V}(t),\mathbf{R}(t)\}$: $\mathbf{V}(t)$ is a spatially defined RV used to provide a suitable probabilistic representation of value; $\mathbf{R}(t)$ is a binary RV defining whether or not a payoff (either positive or negative) is returned;
- $\{\pi(t),z(t),\xi(t)\}$: an *oculomotor state representation* as defined via the binary RV $z(t)$, occurring with probability $\pi(t)$, and determining the choice of motor parameters $\xi(t)$ guiding the actual gaze relocation;
- $\mathcal{D}(t)$: a set of state-dependent statistical decision rules to be applied on a set of candidate new gaze

locations $\mathbf{r}_{new}(t+1)$ distributed according to the posterior distribution on $\mathbf{r}_F(t+1)$.

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

3.2.1 Value and payoff

Following the discussion in Sect. 2, we use the payoff (or reward) as an operational concept for describing the value that the foraging eye gains, under a given task, for landing, after a shift, in $\mathbf{r}_F(t)$. Broadly speaking, it can be conceived as a measure of the "satisfaction of performing well" or the desire to complete a given task. In an object-based setting it amounts to ascribing a value to one or more objects that can be sensed in the FOA region centered in $\mathbf{r}_F(t)$.

In a more formal way, we cast $\mathbf{R}(t)$ as a binary variable, with discrete values of one and zero and we assume that the probability of the *experienced payoff* $\mathbf{R}(t)$, at location $\mathbf{r}(t)$ is described by $P(\mathbf{R}(t)|\mathbf{r}(t), \mathbf{L}(t), \mathbf{O}, \mathbf{T})$. In the vein of [93], payoff magnitude is encoded as the probability $P(\mathbf{R}(t) = 1|\mathbf{r}_F(t), \mathbf{L}(t), \mathbf{O}, \mathbf{T})$, for which we use the shorthand $P(\mathbf{R}(t))$. Under this encoding, a gaze location $\mathbf{r}_F(t)$ associated with large positive payoff would give $P(\mathbf{R}(t) = 1) \simeq 1$. If the state were associated with large negative payoff , $P(\mathbf{R}(t) = 1)$ would fall near zero.

This entails that, if for generality we are going to adopt either positive or negative numerical values for payoff, we need a proper normalisation within the $[0, 1]$ interval to treat such values as probability values. Thus, following [93],

$$P(\mathbf{R}(t)) = 0.5 \left(\frac{R(\mathbf{r}_F(t))}{R_{max}} + 1 \right), \tag{3}$$

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

To compute such probabilities, the *effective payoff*, that is the actual numerical payoff assigned when gazing at $\mathbf{r}_F(t)$, is always computed along the feed stage and as such it is a local payoff [61]: a functional of the probability measure that is positively defined in a region centred on $\mathbf{r}_F(t)$ (e.g., the FOA). Clearly the effective payoff depends on the task $\mathbf{T}$. For instance, in a free viewing task, an implicit reward will be gained by observers that gaze on text or faces, due to their intrinsic attractiveness [20,112]. However in a "look for text" task, a higher payoff will be gained when a text region is recognised within or near the FOA centred on $\mathbf{r}_F(t)$.

We can formalise such intuition, by considering $\mathbf{T}$ as a selector variable [60] that controls the multiplexed conditional probability density $P(\mathbf{R}(t)|\mathbf{r}(t), \mathbf{L}(t), \mathbf{O}, \mathbf{T})$:

$$P(\mathbf{R}(t)|\mathbf{r}_F(t), \mathbf{L}(t), \mathbf{O}, \mathbf{T} = S) = P(\mathbf{R}(t)|\mathbf{r}(t), \mathbf{O}); \tag{4}$$

$$P(\mathbf{R}(t)|\mathbf{r}_F(t), \mathbf{L}(t), \mathbf{O}, \mathbf{T} = FV) = P(\mathbf{R}(t)|\mathbf{r}(t), \mathbf{L}(t)). \tag{5}$$

Eq. 4 is selected when the task is a search task: in this case the effective payoff $R(\mathbf{r}_F(t))$ is a functional of the probability $P(\mathbf{r}_F(t), \mathbf{O})$ of "hitting" an object of class $\mathbf{O}$ while gazing at $\mathbf{r}_F(t)$. Namely, $R(\mathbf{r}_F(t)) = \int_{\mathcal{N}(\mathbf{r}_F(t))} P(\mathbf{r}_F(t), \mathbf{O}) d\mathcal{N}$, where $\mathcal{N}(\mathbf{r}_F(t))$ is a suitable neighborhood centered on current gaze location. This is basically the effective payoff locally computed in terms of a high-resolution object detector. By contrast, in a free-viewing task we compute $R(\mathbf{r}_F(t))$ by taking into account the local landscape of the priority map (Eq. 5). The rationale behind this choice stems from the fact that, although it is clear whether a subject fixates a particular region in a scene, it is not so easy to infer which features are being processed (the difference between looking and seeing [86]). In a search session fixational eye movements are likely to serve the purpose of confirming the identity of a detected object or disambiguating parts of an object; thus, the local use of a classifier/detector working at high-resolution, which is more performant than a weak and lower resolution

localiser as applied in the pre-attentive stage, is a desirable choice [117]. On the other hand, the free-view task is unfortunately very uncontrolled. However, some of the highest correlations between saliency/relevance and fixation are found in free-viewing tasks. This is likely to happen, since in the absence of a specific target, visual saliency coincides with places that are useful for interpreting or remembering the scene [36]. In this case, the choice of computing the local reward as $R(P(\mathbf{r}_F(t), \mathbf{L}(t))$ is a reasonable approach.

The payoff gained at $\mathbf{r}_F(t)$ allows to update the probability distribution of value defined on $\mathbf{V}(t)$, the time-varying spatial map of behaviourally relevant locations over the visual space, so that at each point a task-dependent value is attached. For the specific purposes of this study, we assume a layered representation of value maps, $\{\mathbf{V}_\ell(t)\}_{\ell=1}^{|\mathbf{O}|}$, in particular one for each class of objects that may be relevant for the given task. This is an extension of the scheme proposed by Navalpakkam *et al.* [73], though their study was limited to the use of primary rewards. Each location of $\mathbf{V}_\ell(t)$ represents a binary random variable $\mathbf{v}_\ell(\mathbf{r}, t)$, denoting whether $\mathbf{r}$ is a valuable point ($\mathbf{v}_\ell = 1$) or not ($\mathbf{v}_\ell = 0$).

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

$$P(\mathbf{v}_\ell(\mathbf{r}, t^{'})|\mathbf{R}(t^{'})) = k_V \left(\sum_{t=1}^{t^{'}} E_{P(\mathbf{R})}[\mathbf{R}(t)|\mathcal{N}(\mathbf{r})] + P(\mathbf{v}_\ell(\mathbf{r}, 0)) \right), \quad (6)$$

where k_V is a suitable normalizing constant. Eq. 6 provides an iterative formulation of the recursive computation of the pdf $P(\mathbf{v}_\ell(\mathbf{r}, t)|\mathbf{R}(t), \mathbf{v}_\ell(\mathbf{r}, t-1), \mathbf{T})$.

At time $t = 0$, the ℓ-th density $P(\mathbf{v}_\ell(\mathbf{r}, 0))$ is initialized as a function of $P(\mathbf{L}(t), \mathbf{O} = o|\mathbf{T})$, the object-based map obtained through a pre-attentive rough classification stage (see Sec. 4). The effective value at each point is assigned using Eq. 3. Notice that the value map is different than the priority map $\mathbf{L}$ although at $t = 0$ it might be similar since the distribution $P(\mathbf{L}|\mathbf{I})$ captures the presence of objects (in the sense of shaping an object-based top-down salience map). Indeed, value depends on task and is adapted in time as a function of payoff: for instance in a control task, regions that are likely to contain objects do not loose value in time by always assigning positive rewards, so to be re-fixated; in a "quickly search for all objects", value of the detected object will decrease in time, since reward will be high for the first fixation on the objects and negative for subsequent fixations.

3.2.2 Oculomotor state representation

Once a value setting is supplied, the ultimate problem of gaze relocation is to choose between feeding on local information (*intensive stage* performed through fixational movements) or "flying away" in search of more valuable foraging patches by relocating gaze (*extensive stage* via medium and large saccades) [11]. Notice that we equate fixations with local feeding, since a fixation is not simply the maintenance of the visual gaze on a single location but rather a slow oscillation of the eye (minimum 50 milliseconds duration) within a circumscribed region (typically $0.5^\circ - 2.0^\circ$ degrees of visual angle), [49]; longer displacements stand for saccades. Formally, at any moment t, we index such two states using the binary RV $z(t)$, where $z(t) = 1$ denotes the "feed" state and $z(t) = 0$ the "fly" state.

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

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

$$\pi(t) \propto \exp\left(-\frac{n_s(t)}{\mu(R(\mathbf{r}_F(t)))} \right); \quad (7)$$

To sum, the mean feeding rate, determining the willingness of the forager to continue the feeding stage, is a function of gained payoff, which in turn depends on the given task $\mathbf{T}$. When "biased" parameters $\pi(t)$ have been computed, the oculomotor state can be sampled

as:

$$z(t) \sim Bern(z(t); \pi(t)). \tag{8}$$

3.2.3 Deciding the gaze shift

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

$$\mathbf{r}_{new}(t+1) \sim P(\mathbf{r}_F(t+1)|\mathcal{A}(t), \mathcal{W}(t), \mathbf{r}_F(t)) \tag{9}$$

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

By assuming that the current oculomotor state is $z(t) = k$ and considering the conditional dependencies in the PGM of Fig. 3, Eq. 9 can be reduced to

$$\mathbf{r}_{new}(t+1) \sim P(\mathbf{r}_F(t+1)|\mathcal{O}(t), \boldsymbol{\xi}_k(t), \mathbf{r}_F(t)), \tag{10}$$

where $\boldsymbol{\xi}_k(t)$ are the most likely motor parameters for state $z(t) = k$, from which the angle and the amplitude of the gaze shift can be derived. Parameters $\boldsymbol{\xi}_k(t)$ and candidates $\mathbf{r}_{new}$ are obtained, at the simulation stage, via a stochastic sampling procedure. Indeed, stochastic sampling provides the computational tool to mimic human gaze shift variability (for details, see following Sec. 4 and [12] for an in-depth discussion).

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

$$E\left[R_{\mathbf{r}_{new}}\right] = \sum_{p \in \mathcal{I}_V^k} \mathcal{V}(\mathcal{O}_p(t)) P(\mathcal{O}_p(t)|\mathbf{r}_{new}(t+1), \mathbf{T}). \tag{11}$$

In Eq. 11, the proto-objects $\mathcal{O}_p$ to be considered are those included in the set $\mathcal{I}_V^k$ of most valuable patches sampled from the whole image at time t, whose dimension is $|\mathcal{I}_V^k(t)| = N_V \leq N_p$. In Eq 11, $\mathcal{V}$ is the average value of proto-object $\mathcal{O}_p(t)$ with respect to the probability maps $P(\mathbf{V}_\ell(t)|\mathbf{R}(t))$.

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

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

$$\mathbf{r}_F(t+1) = \arg\max_{\mathbf{r}_{new}} E\left[R_{\mathbf{r}_{new}}\right]. \tag{12}$$

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

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

4 Simulation: gaze shift sampling

Here we provide details of a computational procedure to simulate the main features of the model and also we present some results by elucidating the whole computational process step by step; the corresponding representations that are obtained at the different levels of processing in the simulation are shown in Fig. 4. Following [12], we take the view that the gaze shift $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t+1)$ is a way of sampling the visual landscape $\mathcal{W}(t)$ according to the current oculomotor action setting $\mathcal{A}(t)$ framed by the task $\mathbf{T}$.

Pre-attentive representation We assume that at $t = 0$, when then observer opens his eyes, a quick pre-attentive representation of the scene is made available [83]. To this end the fixation point $\mathbf{r}_F(0)$ is set at the centre of the picture, and the retinal image is simulated by blurring $\mathbf{I}$ through an isotropic Gaussian function centered at $\mathbf{r}_F(t)$, whose variance is taken as the radius of a FOA, $\sigma = |FOA|$, approximately given by $1/8 \min[w, h]$, where $w \times h = |\Omega|$, $|\Omega|$ being the dimension of the image support Ω. This way we obtain the high resolution (HR) foveated image $\widehat{\mathbf{I}}_{HR}(0)$ (Fig. 4, top row, right picture); the foveated HR is mainly exploited to support local fixational movements and small saccades. This is then reduced through a pyramidal decomposition to $\widehat{\mathbf{I}}_{LR}(t)$, a low-resolution (LR) image mainly used during long relocation of the gaze. The foveation process will be updated for every gaze shift involving a large relocation, but not during fixational eye movements.

The LR image is adopted to roughly compute the initial feature likelihood $P(\mathbf{X}|\mathbf{F}, \mathbf{L})$. To such end, for what concerns face objects, we use the Viola-Jones detector by converting the AdaBoost outcome in a probabilistic output [9]. For what concerns textual objects, following [20] we simulate the localizer/detector using the text ground-truth. However, to be more realistic

and compliant with the theoretical model, differently from [20], object likelihood is computed by using the output of Torralba's saliency [104] localised in the bounding box as given by the text region ground-truth. The motivation for this choice is that Torralba's saliency well correlates with text appearance [90] and it can be used as a rough but reliable estimate of its likelihood $P(\mathbf{F}|\mathbf{O} = text)$. Further, the main reason for using a simulated text likelihood estimator (instead of a real one such as in [26]) is that one can exploit *ad-hoc* control of the number of true positive / false positive regions. Having computed these coarse object-based maps it is easy to infer the initial priority map $P(\mathbf{L}|\widehat{\mathbf{I}}_{LR})$ [24] (Fig. 4, second row, left picture).

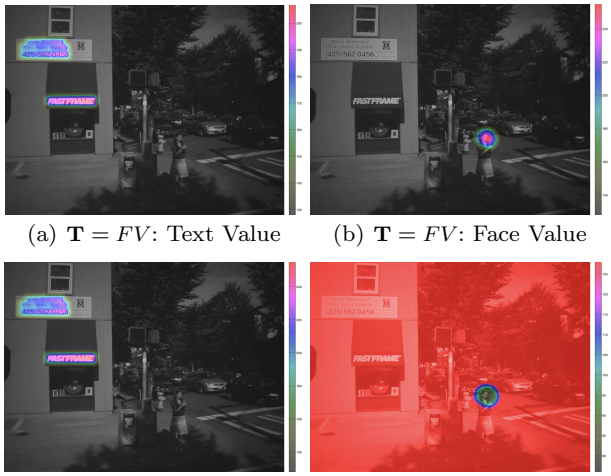
The value probability maps $P(\mathbf{v}_\ell(\mathbf{r}, 0))$ can be initialised as discussed in Sec. 3.2.1. One example, referring to the picture used in Fig. 4 is provided in Fig. 5. More in detail, such initialisation has been obtained through the following steps. At time $t = 0$, the payoffs are set as a function of the task. We used $R_{text} = 50$ and $R_{face} = 100$ for $\mathbf{T} = FV$ (free-view), granting a higher attractiveness to faces with respect to text. For $\mathbf{T} = S$ (searching for text), $R_{text} = 100$ and $R_{face} = -50$. Then, the spatial feature map $P(\mathbf{X}(t)|\mathbf{L}(t), \mathbf{F}(t))$ computed for either $\mathbf{O} = face$ and $\mathbf{O} = text$ provides a pair of object likelihood maps that are used as approximate estimates of the object-based posterior density maps $P(\mathbf{L}(t), \mathbf{O} = face|\mathbf{T})$ (the posterior probability of observing a face object at a spatial location) and $P(\mathbf{L}(t), \mathbf{O} = text|\mathbf{T})$ (the posterior probability of observing a text object). Task $\mathbf{T}$ being assigned, the object maps are multiplied, with the payoff values chosen as above. To this point, the resulting maps are no longer probability maps. Thus, Eq. 3 is applied to each point of the maps for normalising between 0 and 1, and the task dependent value maps are eventually obtained, i.e. $P(\mathbf{V}_{text}(0)|\mathbf{R}(0), \mathbf{T})$, $P(\mathbf{V}_{face}(0)|\mathbf{R}(0), \mathbf{T})$. Such maps are shown in Fig. 5, where, for visualisation purposes, probabilities have been represented through colours. Note that, in order to fairly compare left and right probability maps, each colourbar at the right side of the map represents a colour (probability) range that is specific for that map. For instance the colourbar in Fig. 5(c) depicts the range $[130 = grey, \cdots, 255 = red]$, whilst the colourbar in Fig. 5(c) represents the range $[75 = grey, \cdots, 130 = red]$.

Sparse representation of proto-objects: Similar to the ES model described in [12] we will exploit here a sparse representation of proto-objects. These are conceived in terms of foraging sites around which interest points can be situated (in the ecological metaphor, food items/preys, [12]).

Fig. 4 The main representations that are obtained at the different levels of processing in the simulation (details in the simulation discussion, Sec. 4). In this case the given task $\mathbf{T}$ is a "*Look for text regions*" task. From top to bottom, left to right: the original image $\mathbf{I}$; the foveated image $\widehat{\mathbf{I}}$ obtained by setting the initial FOA $\mathbf{r}_F(0)$ at the centre of the image; the priority map $\mathbf{L}$; selected proto-objects parametrised as ellipses $\theta_p(t)$; the interest points $O(t)$ sampled from proto-objects; the sampling process of candidate FOAs $\mathbf{r}_{new}(t+1)$ (Eq. 17) and the selection of k-th candidate point which maximises the expected reward $E[R_{\mathbf{r}_{new}}]$ (the big circles covers the points within $\mathcal{I}_V^k$); the sampled FOA $\mathbf{r}_F(t+1)$. All maps are depicted at the same resolution (HR) of the original image $\mathbf{I}$ for visualisation purposes. Value map initialisation follows the procedure illustrated in Fig. 5 below

.

At any given time t, the foraging eye perceives a set $\mathcal{O}(t) = \{\mathcal{O}_p(t)\}_{p=1}^{N_P}$ of proto-objects or patches in terms of prey clusters, each patch being characterised by different shape and location. More formally, $\mathcal{O}_p(t) = (O_p(t), \Theta_p(t))$. Here $\Theta_p(t)$ is a parametric description of a patch, while $O_p(t) = \{\mathbf{r}_{i,p}\}_{i=1}^{N_{i,p}}$ is a sparse representation of patch p as the cluster of interest points that can be sampled from it. More precisely, $\Theta_p(t) = (\mathcal{M}_p(t), \theta_p)$. The set $\mathcal{M}_p(t) = \{m_p(\mathbf{r}, t)\}_{\mathbf{r}\in L}$ stands for a map of binary RVs indicating at time t the presence or absence of patch p. The overall map of proto-objects is given by $\mathcal{M}(t) = \bigcup_{p=1}^{N_p} \mathcal{M}_p(t)$. Here, $\mathcal{M}(t)$ is simply drawn from the priority map by deriving a preliminary



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

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

Fig. 5 The initial value probability maps $P(\mathbf{V}_\ell(0)|\mathbf{R}(0), \mathbf{T})$ calculated by weighting, at each spatial location, the estimated object maps (text and face) through the numerical payoff chosen for the given task $\mathbf{T}$ (see text for details). The input image is the one used for the example in Fig. 4. Free view (FV): $P(\mathbf{V}_{text}(0)|\mathbf{R}(0), FV)$ 5(a) and $P(\mathbf{V}_{face}(0)|\mathbf{R}(0), FV)$ 5(b). Search for text (S): $P(\mathbf{V}_{text}(0)|\mathbf{R}(0), S)$ 5(c) and $P(\mathbf{V}_{face}(0)|\mathbf{R}(0), S)$ 5(d). Probabilities, superimposed on the foveated image, have been scaled between $[0, 255]$ and colour coded, red colour denoting high probability, grey colour low probability.

binary map $\widetilde{\mathcal{M}}(t) = \widehat{m}(\mathbf{r}, t)\}_{\mathbf{r}\in L}$, such that $\widehat{m}(\mathbf{r}, t) = 1$ if $P(\mathbf{L}(t)|\widehat{\mathbf{I}}(t)) > T_M$, and $\widehat{m}(\mathbf{r}, t) = 0$ otherwise. The threshold T_M is adaptively set so as to achieve 95% significance level in deciding whether the given priority values are in the extreme tails of the pdf. The procedure is based on the assumption that an informative proto-object is a relatively rare region and thus results in values which are in the tails of $P(\mathbf{L}(t)|\widehat{\mathbf{I}}(t))$. Then, following [111], $\mathcal{M}(t) = \{\mathcal{M}_p(t)\}_{p=1}^{N_P}$ is obtained as $\mathcal{M}_p(t) = \{m_p(\mathbf{r}, t)|\ell(B, \mathbf{r}, t) = p\}_{\mathbf{r}\in L}$, where the function ℓ labels $\widetilde{\mathcal{M}}(t)$ around $\mathbf{r}$ using the classic Rosenfeld and Pfaltz algorithm (implemented in the Matlab bwlabel function). We set the maximum number of patches to $N_P = 8$ to retain the most important patches. The patch map provides the necessary spatial support for a 2D ellipse maximum-likelihood approximation of each patch (see Fig. 4 second row, right picture), whose location and shape are parametrised as $\theta_p = (\mu_p, \Sigma_p)$ for $p = 1, \cdots, N_p$ (see [12] for a formal justification). Next, the procedure generates clusters of interest points, one cluster for each patch p:

$$O_p(t) \sim P(O_p(t)|\theta_p(t), \mathcal{M}_p(t) = 1, \mathbf{L}(t)). \quad (13)$$

By assuming a Gaussian distribution centered on the patch, Eq. (13) can be further specified as [12]:

$$\mathbf{r}_{i,p} \sim \mathcal{N}(\mathbf{r}_p; \mu_p(t), \Sigma_p(t)), i = 1, \cdots, N_{i,p}. \quad (14)$$

We set $N_s = 50$ the maximum number of interest points and for each patch p, and we sample $\{\mathbf{r}_{i,p}\}_{i=1}^{N_{i,p}}$ from a Gaussian centered on the patch as in (14). The number of interest points per patch is estimated as $N_{i,p} = \lceil N_s \times \frac{A_p}{\sum_p A_p} \rceil$, $A_p = \pi\sigma_{x,p}\sigma_{y,p}$ being the area of patch p. Thus, the set of all interest points characterising the perceived scene can be obtained as $O(t) = \bigcup_{p=1}^{N_p}\{\mathbf{r}_{i,p}(t)\}_{i=1}^{N_{i,p}}$ (Fig. 4, third row, left picture).

Determining the oculomotor action setting: At the end of the proto-object sampling procedure we have at time t the set $\mathcal{O}(t) = \{\mathcal{O}_p(t)\}_{p=1}^{N_P}$ of proto-objects in terms of interest points $O(t)$, each patch being characterised by different shape and location, i.e., by proto-object parameters $\Theta_p(t)$. The first step is to determine the oculomotor state by sampling from the Bernoulli distribution via Eq. 8 with parameters determined by Eq. 7.

Assume that choice $z(t) = k$, with $k = 0, 1$, has been made. This allows to set the actual values of the motor parameters $\eta_k = \{\alpha_k, \beta_k, \gamma_k, \delta_k\}$. These are the parameters of the α-stable distribution $f(\boldsymbol{\xi}_k; \eta_k(t))$, namely, the skewness β (measure of asymmetry), the scale γ (width of the distribution), the location δ and, most important, the characteristic exponent α, or index of the distribution that specifies the asymptotic behavior of the distribution. The α-stable distribution $f(\boldsymbol{\xi}_k; \eta_k(t))$ is then used to sample the stochastic components $\boldsymbol{\xi}_k(t) = \{\xi_{k,1}, \xi_{k,2}\}$ of candidate gaze shifts [12]:

$$\boldsymbol{\xi}_k(t) \sim f(\boldsymbol{\xi}_k; \eta_k(t)) \quad (15)$$

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

Deciding where to look next Having determined the oculomotor action setting $\mathcal{A}(t)$, we can rewrite Eq. 10, that is the sampling of candidate gaze locations for the shift $\mathbf{r}_F(t) \to \mathbf{r}_F(t+1)$ as:

$$\mathbf{r}_{new}(t+1) \sim P(\mathbf{r}_F(t+1)|O(t), \theta(t), \boldsymbol{\xi}_k(t), \mathbf{r}_F(t)), \quad (16)$$

where the distribution on the l.h.s. of Eq. 16 is the oculomotor state transition probability of the shift. The shift is generated according to motor behaviour $z(t) = k$ and thus regulated by parameters η_k conditioned on proto-objects sparsely represented through sampled interest points $O(t)$ and patch parameters $\theta(t)$. We sample $\mathbf{r}_{new}(t+1)$ by making explicit the stochastic dynamics behind the process. To this end we exploit the Euler-Maruyama discretisation of a Langevin-type stochastic

differential equation (see [12] for a formal derivation):

$$\mathbf{r}_F(t_{n+1}) \approx \mathbf{r}_F(t_n) - \sum_{p\in\mathcal{I}_V^k}\sum_{i\in\mathcal{I}_p}(\mathbf{r}_F(t_n)-\mathbf{r}_p(t_n))\tau + \gamma_k \mathbb{I}\tau^{1/\alpha_k}\boldsymbol{\xi}_k. \quad (17)$$

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

The term $\gamma_k \mathbb{I}\tau^{1/\alpha_k}\boldsymbol{\xi}_k$ is the stochastic component which determines amplitude and orientation of the candidate gaze shift [12]. The symbol $\mathbb{I}$ denotes the 2×2 identity matrix and γ_k the width of the α-stable distribution from which $\boldsymbol{\xi}_k$ is sampled (Eq. 15). Notice that, due to the feed/fly switching of index $k = z(t)$ in Eq. 17, this random walk is a mixture of Lévy (large relocation) and nearly-Gaussian (local exploration) displacements.

Thus, Eq. 17 provides the explicit procedure for sampling candidate gaze shifts $\mathbf{r}_F(t) \rightarrow \mathbf{r}_{new}(t+1)$. Assume we sample N_{new} such candidates, as shown in Fig. 4, third row, right picture. Then the decision to saccade is taken in order to maximise the expected reward of having valuable interest points in the neighbourhood of the candidate shift (represented in the same picture as a wide yellow circle). This can be obtained by writing Eq. 11 as

$$E\left[R_{\mathbf{r}_{new}}\right] = \sum_{p\in\mathcal{I}_V^k}\sum_{i\in\mathcal{I}_p}\mathcal{V}(\mathbf{r}_{i,p}(t))\mathcal{N}(\mathbf{r}_{i,p}(t)|\mathbf{r}_{new}(t+1),\Sigma_s). \quad (18)$$

Finally, the actual gaze shift is obtained through Eq. 12 (Fig. 4, bottom picture)

Recall from Secs. 3.2.3 and 3.2.1 that in the feeding state we have to compute the effective reward. In particular, if the task is a search task, we stated that the effective reward $R\left(P(\mathbf{r}_F(t),\mathbf{O})\right)$ should be computed through a "high resolution" detector/classifier. To such end, if the object to look for is a face we use the probabilistic version of the Viola-Jones detector, but working on the HR image (which entails higher precision); if we are searching for text, we straightforwardly use the HR text ground-truth as a "perfect classifier" (oracle). To complete the picture, at each shift the IOR is simulated on the priority map by applying an inverse Normal suppression function at $\mathbf{r}_F(t)$, as in [98].

All parameters of the model have been tuned by using a subset of 50 images from the Microsoft dataset and related eye tracking data (see Secs. 5.1, 5.3).

Finally, in order to get a better understanding of the inner workings of the model, we show an example where we successively switch off the different control levels. Results are shown in Fig. 6. The top row presents two scan paths obtained assigning a "*Look for text*" task (left picture) and simulating a "*Look for people*" task (right); at this level the simulation of the model is working in full mode. The central row presents results obtained when no task is given and control by value and payoff is inhibited. The left picture shows the priority map after the first central fixation. In this case, $\mathcal{W}$ relies entirely on the priority representation and the on proto-objects that can be sampled from it; also, the prior probability of objects given the task, $P(\mathbf{O}|\mathbf{T})$, is taken as a uniform distribution and hence the contribution by early salience becomes stronger. The forager's willingness to feed or to fly $\mu(R(\mathbf{r}_F(t)))$ (Eq. 7) is set to a constant, and the decision rule in Eq. 18 is simplified by letting $\mathcal{V}(\mathbf{r}_{i,p}(t)) = \mathbf{r}_{i,p}(t)$, that is $\mathcal{V}$ is to be considered an identity function, since value plays no role at this stage. The right picture on the same row depicts one simulated scan path where the central bias effect of the foveated priority map is readily apparent. The bottom row shows the simulation of the model when no object information is available, thus $P(\mathbf{F}(t)|\mathbf{O}) = P(\mathbf{F}(t))$ and the gaze shift process (right) only nourishes on early salience yet modulated by foveation (left).

5 Experimental work

The experimental work aimed at confronting the scan paths produced by the model with those obtained from either eye-tracked human subjects using data from a publicly available dataset or performing a new eye-tracking experiment on a public dataset of complex urban pictures (see Sec 5.1). Such comparison was qualitative in terms of observable scan paths, but also quantitative in terms of statistical similarity of oculomotor behaviour. To the latter end, gaze shift amplitude distributions of human observers were compared to those obtained by running the simulation. Indeed, the study of the amplitude distribution [100,102], and in particular of the corresponding complementary cumulative distribution function, is the standard convention in the literature [12]. Further, in the specific case of the "*Look for text*" task, we also analysed the discriminability performance of simulated scan paths in terms of average True Positive Rate and False Positive Rate. This in order to provide quantitative results concerning semantic aspects that the search task brings in. For all quantitative assessments we used as a baseline control model, the Itti et *et. al* model [54] as implemented in the lat-

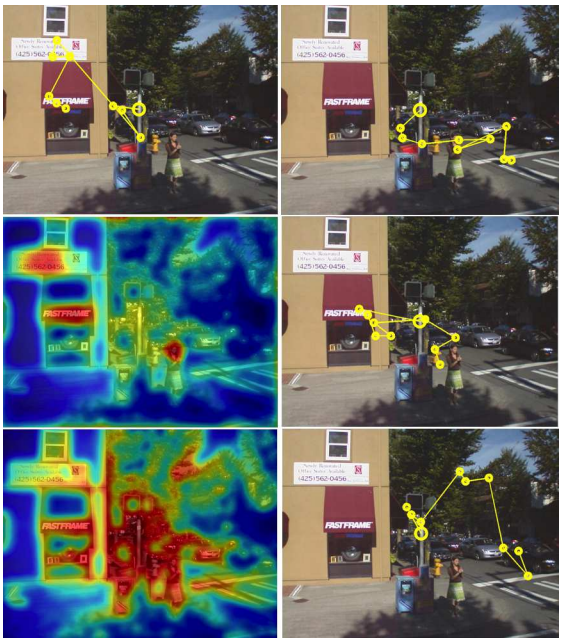


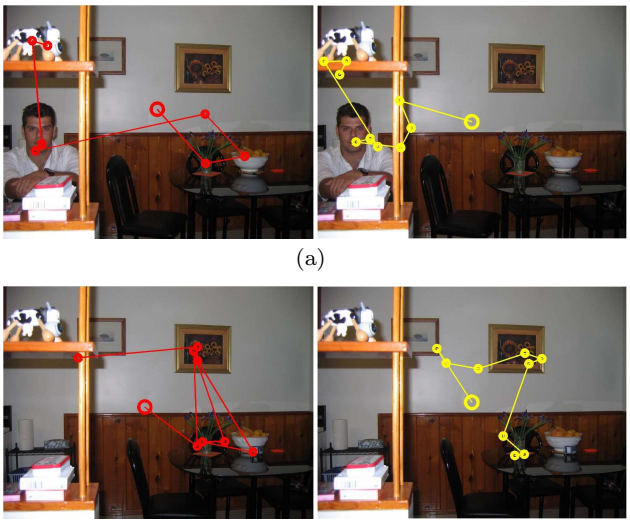
Fig. 6 Inhibition of levels of representation and control. Top row: scan path generated when the given task **T** is "*Look for text*", similarly to Fig. 4 (left); scan path generated when the model simulates a "*Look for people*" task (right). Middle row, no task and value assigned, but object likelihood is still computed: the foveated priority map **L** (left, red colour coding for high priority locations, blue for low priority) and one generated scan path (right). Bottom row: when the object likelihood is not computed, the priority map collapses to a classic early salience but modulated by foveation (left); a corresponding scan path (right). All maps are depicted at the same resolution (HR) of the original image **I** for visualization purposes

est version of the saliency tool box downloaded from http://www.saliencytoolbox.net

5.1 Datasets

Cerf's Fixations in FAces dataset. The dataset is downloadable at `http://www.fifadb.com/`. This dataset contains Faces a subset of 229 images (1024×768 pixels) showing frontal faces in various sizes, locations, skin colours, races, etc. Each image has a corresponding background image with no faces for comparison. The data include the fixations recorded via eye-tracking of 8 subjects (see [20] for details). In addition to fixation data, an annotation of the entire dataset is provided, where the location and labelling of faces in images are given.

Epshtein's Microsoft dataset. For specifically assessing the behaviour of the model on text objects in natural scenes, a publicly available dataset (`http://research.microsoft.com/en-us/um/people/eyalofek`) has been



(a)

(b)

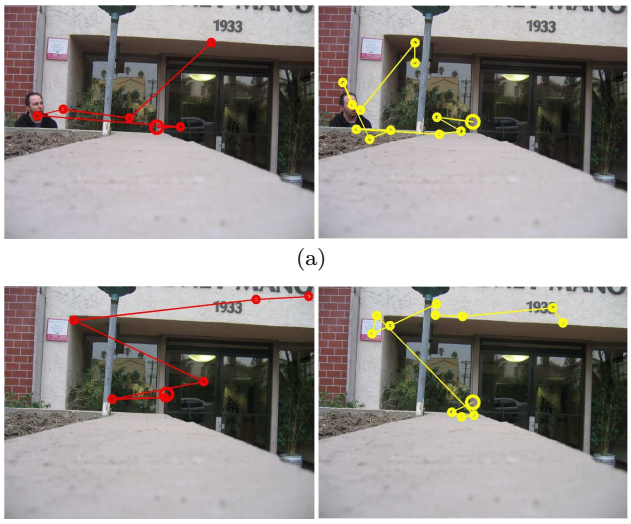
Fig. 7 Scan paths generated while free viewing a picture from Fixations in FAces dataset, when a face is present 7(a) and when the face is removed 7(b). Left (in red colour), scan path obtained eye-tracking a human observer; right, model output (in yellow)

used for testing the behaviour of the model's simulation. This consists of 307 colour street view pictures of sizes ranging from 1360×1024 to 1024×768 pixels. The text content is embedded in the scene in the form of shop names, street signs or advertisements and it is usually not located at the centre of the image, nor covering a large region of the image, so as to make the localisation problem more difficult (see [2] for details).

5.2 Experiment 1

The aim of this experiment was to compare the motor behaviour predicted by the model with experimental scan paths from human subjects in free viewing condition ($\mathbf{T} = FV$). For this experiment we used the Fixations in FAces dataset. Pictures contained either faces, or text regions or both. First comparison is qualitative. We generated 20 scan paths for each image and compared them to those exhibited by human observers by choosing most similar scan paths in terms of fixations coordinates, duration, and time occurrence. Some typical results obtained are presented in Figs. 7, 8 showing the ability of the model to mimic observer's oculomotor behaviour.

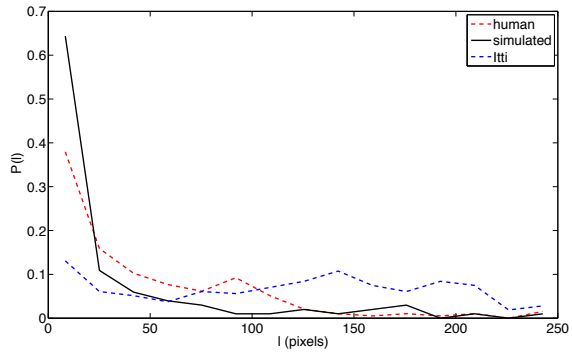
More quantitatively, we studied the empirical distributions of gaze shift amplitudes [102, 100, 12] by analyzing eye-tracking results collected in the dataset To this end, gaze shift samples from all the traces regardless of the observers, are aggregated together and used in



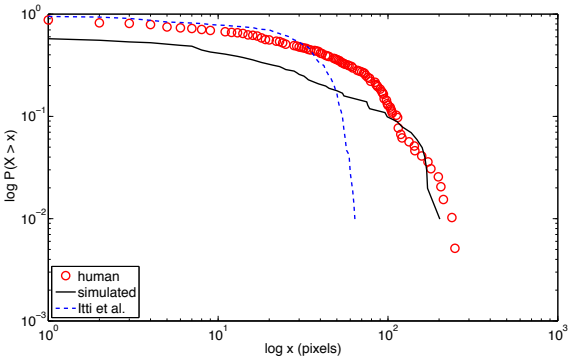
(a)

(b)

Fig. 8 Scan paths generated while free viewing a picture from Fixations in FAces dataset. In 8(a) face and text are both present, whilst in 8(b) the face is removed. Left (in red colour), scan path obtained eye-tracking a human observer; right, model output (in yellow)



(a) Gaze shift amplitude distribution



(b) CCDF

Fig. 9 Comparing the oculomotor behaviour generated by humans with either the one simulated by the proposed model and by the one of Itti. The comparison is provided in terms of gaze shift amplitudes on the Fixations in FAces dataset . Top panel (9(a)) shows the empirical distributions of gaze shift amplitudes; bottom panel (9(b)) shows the double log-plots of the corresponding CCDFs.

the same distribution. The assumption is that every observer under the same task has the same statistical "mobility tendency" in terms of gaze shifts; then this "aggregation" is reasonable because every trace obtained from the same image is subject to the same or similar visual constraints. The same technique is used in other studies of Levy walks (e.g., [84]) but also in eye-tracking experiments [100]. For a more precise description of the tail behaviour, i.e. the laws governing the probability of large shifts, the upper tail of the distribution of the gaze shift magnitude X has also been considered. This can be defined as $\overline{F}(x) = P(X > x) = 1 - F(x)$, where F is the cumulative distribution function (CDF). Consideration of the upper tail, or complementary CDF (CCDF) of jump lengths is the standard convention in the literature. To introduce a control condition, separate simulations were run for virtual observers viewing the same set of images as the human observers, either by using our model and a baseline control model, namely the Itti et *et. al* model [54]. For each image, the virtual observer made the same number of simulated saccades as the human observer had on that scene. Such results are illustrated in Fig. 9

It can be noticed that Itti *et al.* model does not show the characteristic positively skewed distribution of gaze shift amplitudes exhibited by human scan paths and captured by the proposed model. Differences in gaze shift statistics can be easily appreciated from the CCDF plot (Fig. 9(b)), as regards the tail behaviour of the distribution. These results are consistent with results presented by Tatler *et al.* [100].

Given the empirical distributions of eye-tracked and simulation gaze shifts on the dataset, the fit between the two is basically assessed via the two-sample Kolmogorov-Smirnov (K-S) test, which is very sensitive in detecting even a minuscule difference between two populations of data. We also provide results from the standard Mann-Whitney U (MWU) test, to assess the null hypothesis that two samples have the same median (central tendency). All tests are performed at the level of significance $\alpha = 0.05$ and repeated for ten model simulation trials. According to the K-S test, the simulated distribution resulted no significantly different from the human one for 70% of cases (average value for all trials). MWU assessed the same central tendency for 92% of cases. The control model always fails both tests.

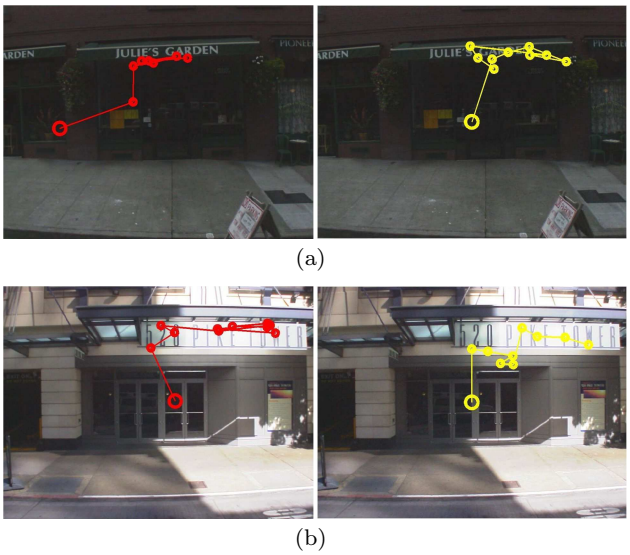
5.3 Experiment 2

For this experiment we used the Microsoft dataset, which comprises images more complex than those in the Fixations in FAces dataset. The use of this dataset offers the advantage of having at hand ground-truth for text regions, but, unfortunately, no eye-tracking data is available, since this dataset is mostly adopted for "text-in-the-wild" detection/classification contests.

Thus, in this case, an eye-tracking assessment have been conducted using a video-based SMI RED eye tracker (SensoMotoric Instruments, Teltow, Germany) at a sampling rate of 120Hz., with automatic head movement compensation (tracking range, 40×30 cm at 70 cm distance). The infrared video-based system has an instrument spatial resolution of 0.03° and an absolute gaze position accuracy of up to 0.4°. The experiment took place in a dimly lit room in the Computer Vision Center in Barcelona. Two groups of six naive adults (3 women and 3 men, composing the first group, 2 women and 4 men for the second group, range 25-44 years, mean 32 years) participated in the experiment. All participants were native speakers of Spanish and had normal or corrected to normal vision. Subjects were seated in a contact-free setup, 70 cm in front of a 22-inch LCD monitor (60 Hz refresh rate, 58.18 dpi). Stimulus resolution was 1024×768 pixels at both sites and subtended approximately a visual angle of $36.6°(w) \times 27.4°(h)$. A 9-point calibration of the eye tracker was carried out at the onset of every trial. Each subject was asked to look at pictures presented on the monitor. Two tasks were considered. A search task, $\mathbf{T} = S$, formulated in terms of "*Look for text regions within the pictured scene*" was assigned to the first group; a free-view task, $\mathbf{T} = FV$, formulated as a generic "*Guess the city from the pictured scene*," so as to motivate the participants, was given to the second group. Pictures were presented in randomized order and each picture was shown for 5 seconds. Stimulus luminance was linear in pixel values.

Qualitative comparison was performed as in Experiment 1. Some examples representative of results obtained for the $\mathbf{T} = S$ case are provided in Figs 10, 11

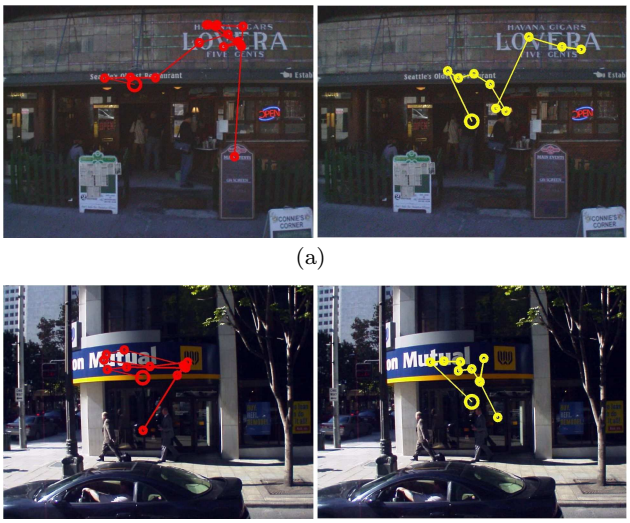
It is worth noting that in the Microsoft dataset the vast majority of images contains text regions as the most semantically relevant objects appearing within the imaged scene. This is reflected in the cumulative statistics of shift amplitudes, which result to be fairly similar for both tasks, as it can be appreciated at a glance from Figs. 13 and 14 below. This was somehow expected, being text attractive even in free-viewing [20,112]. Nevertheless, there are cases that bear a specific interest. For instance, we show one such example in Fig. 12. This can be considered as the "dual" of the example provided in



(a)

(b)

Fig. 10 Scan paths generated under the "*Look for text regions*" task for pictures from the Microsoft dataset, where text is the main semantic object class. Left (in red colour), scan path obtained eye-tracking a human observer; right, model output



(a)

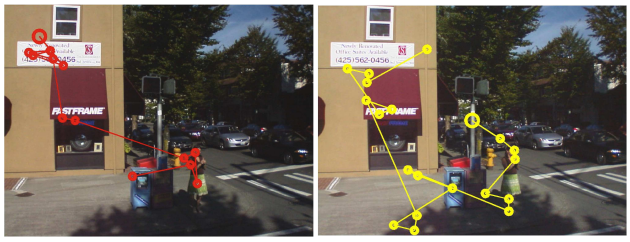
(b)

Fig. 11 Scan paths generated under the *look for text* for pictures from Microsoft dataset when other semantic objects (faces, people) are embedded in the picture together with text. Left (in red colour), scan path obtained eye-tracking a human observer; right, model output

Fig. 8. In that case, under the same task ($\mathbf{T} = FV$), one class of objects was removed. Here, both classes of objects ($\mathbf{O} = face$ and $\mathbf{O} = text$) are retained, but the task is switched from $\mathbf{T} = S$ (12(a)) to $\mathbf{T} = FV$ (12(b)).

It can be noted that for $\mathbf{T} = S$ (12(a), left), the girl is treated as a "distractor" by the human observer, whilst for $\mathbf{T} = FV$ (12(b)) it is competing for attract-

(a) "*Look for text* "



(b) "*Guess the city*"

Fig. 12 Scan paths generated under the "*Look for text* " task, for a picture where other semantic objects (faces, people) are embedded in the picture 12(a) and under the "*Guess the city*" task, 12(b). Left (in red colourrealis), scan path obtained eye-tracking a human observer; right, model output (yellow)

ing gaze though being less visible and physically salient with respect to text regions in the scene (12(b)). The model achieves a similar behaviour by the different assignment of value in either task (cfr. Fig. 5).

In order to provide quantitative results concerning semantic aspects that the "*Look for text* " task brings in, we have performed the following analysis. Since the Microsoft dataset includes the maps of text objects located in each image we can compute the ground-truth binary text map $\mathcal{TM}$ with $\mathcal{TM}(x,y) = 1$ for pixels (x,y) belonging to target objects, $\mathcal{TM}(x,y) = 0$ for points outside text regions. Given the s-th scan path on the same image, we obtain the binary fixation map $\mathcal{FM}_s$ by considering the first 10 fixations of s and by setting to 1 points within the circular region defining around each fixation point, and to 0 points outside such areas. For what concerns the radius of each fixational region, we set $\varphi = 2°$ of visual angle. The size of this "functional fovea" is slightly larger than the 2° window spanned by a fixational eye movement [49,97] and corresponds to the $7° - 8°$ window that can be searched effectively in one fixation [41]. Yet, it is smaller than the conservative estimate by Shioiri and Ikeda, who define 10° of visual angle the maximal window over which high-resolution pictorial information can be extracted [91]. By taking into account the experimental viewing conditions adopted to record the eye-tracking data (viewing distance $v_d = 70$ *cm*, screen resolution $s_r = 58.18$ *dpi*), the radius φ of region can be calculated in pixel units

Table 1 TPR, FPR and d' for observers and virtual observers simulated by the proposed model and by the control model

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

as

$$r_{fix} = \varphi \frac{1}{2\tan^{-1}\left(\frac{1}{2v_d}\right)} \frac{\pi}{180} \frac{s_r}{2.54} \quad (pxl). \tag{19}$$

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

$$TP_s = \mathcal{TM} \bigcap \mathcal{FM}_s, \quad FP_s = \mathcal{TM}^{\complement} \bigcap \mathcal{FM}_s, \tag{20}$$

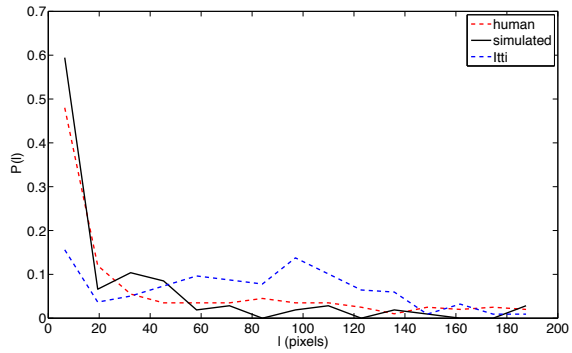
where $\mathcal{TM}^{\complement}$ is the complement of the binary map $\mathcal{TM}$. Then, the average TPR_s and FPR_s are calculated taking into account all the scan paths generated within each group of observers: human, model and control model. The final total averages TPR and FPR computed on all the images of the dataset for each group are reported in Table 1, where the performance of the proposed model can be compared with human and control model performance. As previously, the Itti *et al.* model was used as a baseline control model.

It can be seen from Table 1 that the average sensitivity (TPR) - in our case the average proportion of actual positives (pixels belonging to text regions) that have been correctly spotted within the first 10 fixations - is similar in both human and model generated scan paths, while the control model exhibits a lower sensitivity. Analogously, humans and model are close in terms of specificity ($1-FPR$), at variance with the control model, which is characterized by marginally lower specificity. These results are statistically significant as it can be seen by measuring the difference between the spotting error rate of human observers and the error rate of a model m (either the proposed or the control model). This way, the statistic $z_{obs,m} = (p_{obs} -$

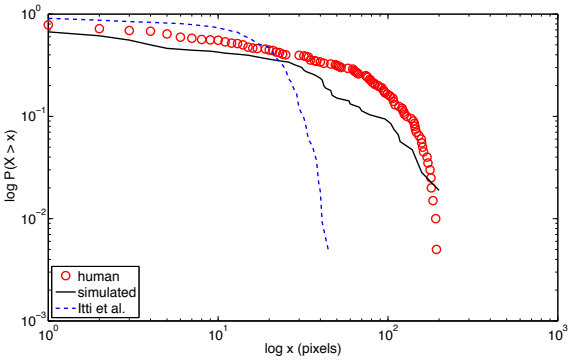
$p_m)/\sqrt{2p(1-p)/n}$ [92] is obtained, with $p = (p_{obs} + p_m)/2$, $n = |N|+|T|$, and where p_{obs} and p_m are the proportions of test samples (pixels) incorrectly spotted by observers and the model m respectively. The statistic has a standard normal distribution [92], and the null hypothesis that human subjects and the model have the same error rate cannot be rejected ($|z_{obs,model}| = 0.07 < Z_{0.975} = 1.96$, two-sided test, $p = 0.94$, significance level $\alpha = 0.05$); conversely, the difference between the control model and humans is remarkable ($|z_{obs,control}| = 70.5 > Z_{0.975}$, $p < 0.001$). The same conclusion is achieved via McNemar's chi-square test [34], with Yates' correction ($p = 0.97$ and $p < 0.001$, respectively, $\alpha = 0.05$).

Similar results are obtained by computing, as index of performance, the discriminability d' (cfr. Table 1), which summarises the capability of the scan path to separate text objects and non text regions, regardless of the statistical decision criterion. This index was calculated as $Z_{TPR} - Z_{FPR}$, where Z_{TPR} is the z-transformed TPR and Z_{FPR} is the z-transformed FPR.

Eventually, as in Experiment 1, Figs. 14 and 13 compare amplitude distributions for either $\mathbf{T} = S$ to $\mathbf{T} = FV$, respectively. Under "*Look for text regions*" task, by performing the K-S test as in the previous experiment, the simulated distribution resulted no significantly different from the human one for an average 79% of cases. MWU assessed the same central tendency for 89% of cases. For the task "*Guess the city*", the K-S test found no significant differences between the two distributions 89% of cases. MWU assessed the same central tendency 96% of cases.



(a) Gaze shift amplitude distribution



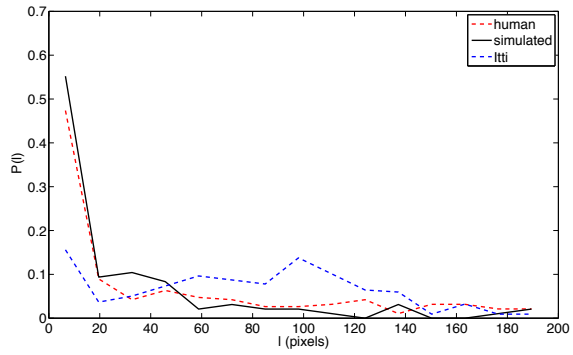
(b) CCDF

Fig. 13 Comparing the oculomotor behaviour generated by humans and simulated by the model on the Microsoft dataset in terms of gaze shift amplitudes. The task was "*Look for text regions*". Top panel (13(a)) compares the empirical distribution of gaze shift amplitudes; bottom panel (13(b)) shows the double log-plot of the corresponding CCDF.

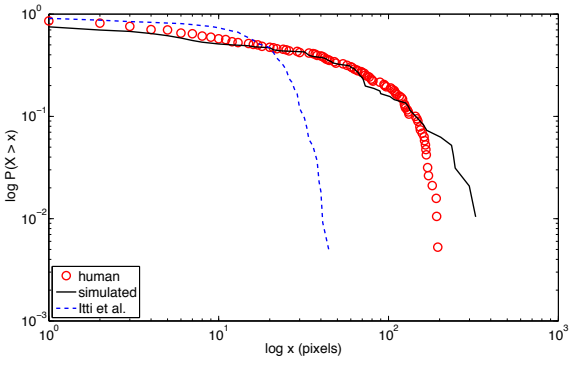
6 Discussion and final remarks

We have presented an integrated computational model of eye guidance for task-dependent attention deployment to objects in natural pictures. To the best of our knowledge, the model is novel in proposing a unified framework that i) accounts for task-dependent visual attention on semantically rich natural images by using different levels of representation, beyond the baseline salience maps; ii) simulates gaze shifts that exhibit statistical properties close to those of eye-tracked subjects, by extending previous approaches proposed in the literature, [12,11] that addressed the intrinsic stochasticity of gaze shifts.

For what concerns the first issue, the proposed model can cope with eye guidance both under a search task, an issue which has been taken into account by some models [115,82,72,117], and under a generic picture viewing task, which has been typically accounted for by salience/relevance-based models (either bottom-up [54] or top-down biased [98,23]). The key to such integration is that, different from those models, we have considered the generation of a scan path as the interplay among several levels of representation and control that goes beyond the classic debate bottom-up vs. top-down, but brings payoff, value and motor representations into the game. We believe that, although this broad and flexible approach also creates new theoretical and computational challenges, this very breadth is an important issue to address. In fact, to succeed in complex environments we must act in a flexible manner as appropriate for a given task, which suggests that a stage of visual selection can be distinct from that of saccade motor selection. For instance, the priority map may encode signals of visual selection that are not eventually captured by current action based decision module. Differently from methods using purely visual top-down modules, the biases provided do not amount directly to motor command, and action related areas may also block or supplement its signal as required in a given task. This integration of different levels of rep-



(a) Gaze shift amplitude distribution



(b) CCDF

Fig. 14 Comparing the oculomotor behaviour generated by humans and simulated by the model on the Microsoft dataset in terms of gaze shift amplitudes. The task was "*Guess the city*". Top panel (14(a)) compares the empirical distribution of gaze shift amplitudes; bottom panel (14(b)) shows the double log-plot of the corresponding CCDF.

resentation and control is important to define some issues that remain elusive if considered with respect to a single level or locus. One such example is the IOR. Depending on the task, different variants of IOR exist [113]. In our model, the priority map explicitly uses IOR in a classic way, by suppressing the response at the currently attended location. However, a reduction or an enhancement in the reward likelihood can modulate the IOR at the priority map level. In general, this multiple level interaction can be a way of framing the discussion surrounding a functional interpretation of IOR (that is fostering or not optimal foraging behaviour, see [113]).

More generally, the use of value and payoff can provide a suitable bridge to explain gaze behaviour that, even in the absence of given task seems to be driven by internal *motivational salience*, which in pathological conditions could be generated by a disruption of biological reward systems [19]. Visual scan path analyses provide important information about attention allocation and attention shifting during visual exploration of social situations characterised by both cognitive complexity and emotional content or even strain. On the one hand, the approach proposed here paves the way to the effective exploitation of computational attention models in the emerging domain of social signal processing [109], and, more broadly, to cope with the problem the affective modulation of the visual processing stream [77,78] with the aim of closing the gap between emotion and cognition [42]

As regards the second issue, scan path variability, the model attempts at filling a gap in the current computational literature (cfr., [14]). The majority of models in computational vision basically resort to deterministic mechanisms to realise gaze shifts, and this, paradoxically, has been the main route to model saccades the most random type of gaze shift [100]. Hence, if the same saliency map is provided as input, they will basically generate the same scan path; further, disregard of motor strategies and tendencies that characterise gaze shift programming results in distributions of gaze shift amplitudes different from those that can be derived from eye-tracking experiments. We have presented in Section 5 examples showing that the overall distributions of human and model generated shifts are close in their statistics, see Figs. 9, 13 and 14. The core of such strategy actually relies upon a mixture of α-stable motions modulated by the different visuomotor levels of control participating to the action-perception loop. The composition of random walks in terms of a mixture of α-stable components allows to treat different types of eyes movement (saccades, fixational movements) within the same framework and makes a step towards the unified modelling of different kinds of gaze shifts. The latter is a research trend that is recently gaining currency in the eye movement realm [74]. For instance, when Eq. (17) is exploited for within-patch exploration, it generates a first-order Markov process, which is compatible with most recent findings [7]. Notice that this approach may be exploited for a principled modelling of individual differences and departure from optimality [71] since providing cues for defining the informal notion of scan path idiosyncrasy in terms of individual gaze shift distribution parameters. The latter represents a crucial issue both for theory [89,102] and applications [62]. For instance, the study by Sprenger *et al.* [95], concerning patients with schizophrenia, has shown that show that alterations such as restricted free visual exploration were present in patients independently of cognitive complexity, emotional strain or physical properties of visual cues implying that they represent a rather general deficit, which may be accounted for in terms of group specific oculomotor bias or scanning strategy. Beside theoretical relevance for modelling human behaviour, a stochastic attention selection mechanism can

be an advantage in computer vision and action learning tasks [67,70].

Clearly, there are some limitations of the model in its present version. We do not consider here time-varying or multiple task assignment, which may be important in real world behaviours. Also, we barely touched the level of neural implementation. However, in this respect, the model is agnostic about whether or not probabilistic computations can be neurally implemented (see the review by Knill and Pouget [58]). This is an intriguing but intricate debate. For instance, Heinke and Humphreys [44] raised the interesting point of using differential equations the exhibits chaotic behaviour to account for noise and recently Churchland and Abbot [25] argued that randomness in neuronal firing rates and spike timing could arise from a network built of deterministic neurons with balanced excitation and inhibition. Further, to make the broad integration behind the model feasible, we have focused on the core issues, providing some black-box or simulated implementations for other components. For instance, for the text localisation/detection task we rely on simulated detectors both for the pre-attentive coarse grained localisation and for the fine-grained detection/recognition. In a preliminary work using a simpler version of the model presented here [26] we have experimented with a text localiser component based on a Relevance Vector Machine classifier applied to "gist" texture features *á la* Torralba [104] both at a coarse and at a high resolution level. However, "textual objects" are a difficult task as opposed to faces for which, at least, efficient and effective face detectors do exist [110], if one is not concerned with the biological plausibility of the algorithm. Actually, our current research work is indeed addressed at verifying the suitability of our model in a difficult practical problem such as text localisation and detection "in the wild", in order to overcome present limitations of attentive-based approaches proposed within such realm [26]. To this end, we are adapting the model to handle time-varying images, and we are performing mobile eye-tracking experiments outside the lab, in complex urban environment. Another limitation, which is conceptually more important than the previous one, is that using value and payoff calls for adopting learning procedures that could be at hand with such information and could be exploited, in the case of a search task, for priming the guidance process [94]. However, it is clear that when dealing with restricted real-world tasks (e.g., crossing a road or making a tea cup) the learning stage can be effectively stated; what has to be learned in the task of searching in a dataset of mostly unrelated pictures of natural scenes is less evident. Treatment of these topics is deferred to a future study.

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

Acknowledgments

The authors are grateful to the Referees and the Associate Editor, for their enlightening and valuable comments that have greatly improved the quality and clarity of an earlier version of this paper. This work was partially supported by the Spanish projects TIN2011-24631, TIN2009-14633-C03-03, CONSOLIDER INGENIO CSD2007-00018 and the fellowships RYC-2009-05031 and 2009FIB00020. With support from the Commission for Universities and Research Department for Innovation, Universities and Enterprise of the Generalitat of Catalonia and the European Social Fund.

References

1. Anderson, B.A.: A value-driven mechanism of attentional selection. Journal of vision **13**(3) (2013)
2. B. Epshtein E. Ofek, Y.W.: Detecting text in natural scenes with stroke width transform. In: 2010 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2963–2970 (2010)
3. Bahill, A.T., Adler, D., Stark, L.: Most naturally occurring human saccades have magnitudes of 15 degrees or less. Investigative Ophthalmology & Visual Science **14**(6), 468–469 (1975)
4. Bartumeus, F., da Luz, M.G.E., Viswanathan, G., Catalan, J.: Animal search strategies: a quantitative random-walk analysis. Ecology **86**(11), 3078–3087 (2005)
5. van Beers, R.: The sources of variability in saccadic eye movements. The Journal of Neuroscience **27**(33), 8757–8770 (2007)
6. Berridge, K.C., Robinson, T.E.: Parsing reward. Trends in neurosciences **26**(9), 507–513 (2003)
7. Bettenbuhl, M., Rusconi, M., Engbert, R., Holschneider, M.: Bayesian selection of markov models for symbol sequences: Application to microsaccadic eye movements. PLoS ONE **7**(9), e43,388 (2012)
8. Boccignone, G.: Nonparametric bayesian attentive video analysis. In: Proc. 19th International Conference on Pattern Recognition, ICPR 2008, pp. 1–4. IEEE Press (2008)
9. Boccignone, G., Campadelli, P., Ferrari, A., Lipori, G.: Boosted tracking in video. Signal Processing Letters, IEEE **17**(2), 129–132 (2010)
10. Boccignone, G., Ferraro, M.: Modelling gaze shift as a constrained random walk. Physica A: Statistical Mechanics and its Applications **331**(1-2), 207–218 (2004)

11. Boccignone, G., Ferraro, M.: Feed and fly control of visual scanpaths for foveation image processing. annals of telecommunications-annales des télécommunications **68**(3-4), 201–217 (2013)
12. Boccignone, G., Ferraro, M.: Ecological sampling of gaze shifts. IEEE Trans. on Cybernetics **44**(2), 266–279 (2014)
13. Boccignone, G., Marcelli, A., Napoletano, P., Di Fiore, G., Iacovoni, G., Morsa, S.: Bayesian integration of face and low-level cues for foveated video coding. IEEE Transactions on Circuits and Systems for Video Technology **18**(12), 1727–1740 (2008)
14. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 185–207 (2013)
15. Borji, A., Sihite, D.N., Itti, L.: An object-based bayesian framework for top-down visual attention. In: Twenty-Sixth AAAI Conference on Artificial Intelligence (2012)
16. Brockmann, D., Geisel, T.: The ecology of gaze shifts. Neurocomputing **32**(1), 643–650 (2000)
17. Bundesen, C.: A computational theory of visual attention. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **353**(1373), 1271–1281 (1998)
18. Canosa, R.: Real-world vision: Selective perception and task. ACM Transactions on Applied Perception **6**(2), 11 (2009)
19. Castellanos, E.H., Charboneau, E., Dietrich, M.S., Park, S., Bradley, B.P., Mogg, K., Cowan, R.L.: Obese adults have visual attention bias for food cue images: evidence for altered reward system function. International Journal of Obesity **33**(9), 1063–1073 (2009)
20. Cerf, M., Frady, E., Koch, C.: Faces and text attract gaze independent of the task: Experimental data and computer model. Journal of Vision **9**(12) (2009)
21. Cerf, M., Harel, J., Einhäuser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. Advances in neural information processing systems **20** (2008)
22. Chambers, J., Mallows, C., Stuck, B.: A method for simulating stable random variables. J. Am. Stat. Ass. **71**(354), 340–344 (1976)
23. Chernyak, D.A., Stark, L.W.: Top–down guided eye movements. IEEE Trans. Systems Man Cybernetics - B **31**, 514–522 (2001)
24. Chikkerur, S., Serre, T., Tan, C., Poggio, T.: What and where: A bayesian inference theory of attention. Vision research **50**(22), 2233–2247 (2010)
25. Churchland, M.M., Abbott, L.: Two layers of neural variability. Nature neuroscience **15**(11), 1472–1474 (2012)
26. Clavelli, A., Karatzas, D., Llados, J., Ferraro, M., Boccignone, G.: Towards modelling an attention-based text localization process. In: J. Sanches, L. Micó, J. Cardoso (eds.) Pattern Recognition and Image Analysis, *Lecture Notes in Computer Science*, vol. 7887, pp. 296–303. Springer Berlin Heidelberg (2013)
27. deCroon, G., Postma, E., van den Herik, H.J.: Adaptive Gaze Control for Object Detection. Cognitive Computation **3**, 264–278 (2011)
28. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Annual review of neuroscience **18**(1), 193–222 (1995)
29. Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., Holmqvist, K.: It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. Behavior research methods **44**(4), 1079–1100 (2012)
30. Dorr, M., Martinetz, T., Gegenfurtner, K., Barth, E.: Variability of eye movements when viewing dynamic natural scenes. Journal of Vision **10**(10) (2010)
31. Einhäuser, W., Rutishauser, U., Koch, C.: Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. Journal of Vision **8**(2) (2008)
32. Einhäuser, W., Spain, M., Perona, P.: Objects predict fixations better than early saliency. Journal of Vision **8**(14) (2008). DOI 10.1167/8.14.18. URL `http://www.journalofvision.org/content/8/14/18.abstract`
33. Ellis, S., Stark, L.: Statistical dependency in visual scanning. Human Factors: The Journal of the Human Factors and Ergonomics Society **28**(4), 421–438 (1986)
34. Everitt, B.S.: The analysis of contingency tables, vol. 45, 2nd edn. CRC Press (1992)
35. Feng, G.: Eye movements as time-series random variables: A stochastic model of eye movement control in reading. Cognitive Systems Research **7**(1), 70–95 (2006)
36. Foulsham, T., Teszka, R., Kingstone, A.: Saccade control in natural images is shaped by the information visible at fixation: evidence from asymmetric gaze-contingent windows. Attention, Perception, & Psychophysics **73**(1), 266–283 (2011)
37. Foulsham, T., Underwood, G.: What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition. Journal of Vision **8**(2) (2008)
38. Frintrop, S., Rome, E., Christensen, H.: Computational visual attention systems and their cognitive foundations: A survey. ACM Trans. on Applied Perception **7**(1), 6 (2010)
39. Fuster, J.: Upper processing stages of the perception–action cycle. Trends in cognitive sciences **8**(4), 143–145 (2004)
40. Gottlieb, J., Balan, P.: Attention as a decision in information space. Trends Cognitive Science **14**(6), 240–248 (2010)
41. Greenwood, P., Parasuraman, R.: Scale of attentional focus in visual search. Perception & Psychophysics **61**(5), 837–859 (1999)
42. Gros, C.: Cognition and emotion: perspectives of a closing gap. Cognitive Computation **2**(2), 78–85 (2010)
43. Hacisalihzade, S., Stark, L., Allen, J.: Visual perception and sequences of eye movement fixations: A stochastic modeling approach. IEEE Trans. Syst., Man, Cybern. **22**(3), 474–481 (1992)
44. Heinke, D., Backhaus, A.: Modelling visual search with the selective attention for identification model (vs-saim): a novel explanation for visual search asymmetries. Cognitive computation **3**(1), 185–205 (2011)
45. Heinke, D., Humphreys, G.W.: Attention, spatial representation, and visual neglect: simulating emergent attention and spatial memory in the selective attention for identification model (saim). Psychological review **110**(1), 29 (2003)
46. Heinke, D., Humphreys, G.W.: Computational models of visual selective attention: A review. Connectionist models in cognitive psychology **1**(4), 273–312 (2005)
47. Hikosaka, O., Nakamura, K., Nakahara, H.: Basal ganglia orient eyes to reward. Journal of Neurophysiology **95**(2), 567–584 (2006)
48. Ho Phuoc, T., Guérin-Dugué, A., Guyader, N.: A computational saliency model integrating saccade programming. In: Proc. Int. Conf. on Bio-inspired Systems and Signal Processing, pp. 57–64. Porto, Portugal (2009)

49. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Van de Weijer, J.: Eye tracking: a comprehensive guide to methods and measures. Oxford University Press, Oxford, UK (2011)
50. Horowitz, T., Wolfe, J.: Visual search has no memory. Nature **394**(6693), 575–577 (1998)
51. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: Proceedings CVPR '07, vol. 1, pp. 1–8 (2007)
52. Humphreys, G.W., Muller, H.J.: Search via recursive rejection (serr): A connectionist model of visual search. Cognitive Psychology **25**(1), 43–110 (1993)
53. Ikeda, T., Hikosaka, O.: Reward-dependent gain and bias of visual responses in primate superior colliculus. Neuron **39**(4), 693–700 (2003)
54. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**, 1254–1259 (1998)
55. Keech, T., Resca, L.: Eye movements in active visual search: A computable phenomenological model. Attention, Perception, & Psychophysics **72**(2), 285–307 (2010)
56. Kimura, A., Pang, D., Takeuchi, T., Yamato, J., Kashino, K.: Dynamic markov random fields for stochastic modeling of visual attention. In: Proc. ICPR '08., pp. 1–5. IEEE (2008)
57. Knill, D., Kersten, D., Yuille, A.: Introduction: A bayesian formulation of visual perception. In: D. Knill, W. Richards (eds.) Perception as Bayesian Inference, pp. 1–21. Cambridge University Press (1996)
58. Knill, D.C., Pouget, A.: The bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences **27**(12), 712–719 (2004)
59. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology **4**(4), 219–27 (1985)
60. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press, Cambridge, MA (2009)
61. Krause, A., Guestrin, C.: Optimal value of information in graphical models. Journal of Artificial Intelligence Research **35**, 557–591 (2009)
62. Le Meur, O., Baccino, T., Roumy, A.: Prediction of the inter-observer visual congruency (iovc) and application to image ranking. In: Proc. 19th ACM international conference on Multimedia, pp. 373–382 (2011)
63. Leon, M.I., Shadlen, M.N.: Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. Neuron **24**(2), 415–425 (1999)
64. Logan, G.D.: The code theory of visual attention: an integration of space-based and object-based attention. Psychological review **103**(4), 603 (1996)
65. Marat, S., Rahman, A., Pellerin, D., Guyader, N., Houzet, D.: Improving visual saliency by adding face feature mapand center bias. Cognitive Computation **5**(1), 63–75 (2013)
66. Marr, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W.H. Freeman, New York (1982)
67. Martinez, H., Lungarella, M., Pfeifer, R.: Stochastic Extension to the Attention-Selection System for the iCub. University of Zurich, Tech. Rep (2008)
68. Maunsell, J.H.: Neuronal representations of cognitive state: reward or attention? Trends in cognitive sciences **8**(6), 261–265 (2004)
69. Mozer, M.C.: Early parallel processing in reading: A connectionist approach. Lawrence Erlbaum Associates, Inc (1987)
70. Nagai, Y.: Stability and sensitivity of bottom-up visual attention for dynamic scene analysis. In: Proceedings of the 2009 IEEE/RSJ international conference on Intelligent robots and systems, pp. 5198–5203. IEEE Press (2009)
71. Najemnik, J., Geisler, W.: Optimal eye movement strategies in visual search. Nature **434**(7031), 387–391 (2005)
72. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. Vision research **45**(2), 205–231 (2005)
73. Navalpakkam, V., Koch, C., Rangel, A., Perona, P.: Optimal reward harvesting in complex perceptual environments. Proceedings of the National Academy of Sciences **107**(11), 5232–5237 (2010)
74. Otero-Millan, J., Troncoso, X., Macknik, S., Serrano-Pedraza, I., Martinez-Conde, S.: Saccades and microsaccades during visual fixation, exploration, and search: foundations for a common saccadic generator. Journal of Vision **8**(14) (2008)
75. Over, E., Hooge, I., Vlaskamp, B., Erkelens, C.: Coarse-to-fine eye movement strategy in visual search. Vision Research **47**, 2272–2280 (2007)
76. Palmer, J., Verghese, P., Pavel, M.: The psychophysics of visual search. Vision research **40**(10), 1227–1268 (2000)
77. Pessoa, L.: On the relationship between emotion and cognition. Nature Reviews Neuroscience **9**(2), 148–158 (2008)
78. Pessoa, L., Adolphs, R.: Emotion processing and the amygdala: from a'low road'to'many roads' of evaluating biological significance. Nature Reviews Neuroscience **11**(11), 773–783 (2010)
79. Peterson, M.S., Kramer, A.F., Wang, R.F., Irwin, D.E., McCarley, J.S.: Visual search has memory. Psychological Science **12**(4), 287–292 (2001)
80. Phaf, R.H., Van der Heijden, A., Hudson, P.T.: Slam: A connectionist model for attention in visual selection tasks. Cognitive Psychology **22**(3), 273–341 (1990)
81. Platt, M.L., Glimcher, P.W.: Neural correlates of decision variables in parietal cortex. Nature **400**(6741), 233–238 (1999)
82. Rao, R.P., Zelinsky, G.J., Hayhoe, M.M., Ballard, D.H.: Eye movements in iconic visual search. Vision Research **42**(11), 1447 – 1463 (2002)
83. Rensink, R.: The dynamic representation of scenes. Visual Cognition **1**(3), 17–42 (2000)
84. Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S., Chong, S.: On the levy-walk nature of human mobility. IEEE/ACM Transactions on Networking, **19**(3), 630–643 (2011)
85. Robert, C.: The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer (2007)
86. Rothkopf, C., Ballard, D., Hayhoe, M.: Task and context determine where you look. Journal of Vision **7**(14) (2007)
87. Rutishauser, U., Koch, C.: Probabilistic modeling of eye movement data during conjunction search via feature-based attention. Journal of Vision **7**(6) (2007)
88. Scholl, B.: Objects and attention: the state of the art. Cognition **80**(1-2), 1–46 (2001)
89. Schütz, A., Braun, D., Gegenfurtner, K.: Eye movements and perception: A selective review. Journal of Vision **11**(5) (2011)

90. Shahab, A., Shafait, F., Dengel, A., Uchida, S.: How salient is scene text? In: Proc. 10th IAPR International Workshop onDocument Analysis Systems (DAS, 2012), pp. 317–321. IEEE (2012)
91. Shioiri, S., Ikeda, M.: Useful resolution for picture perception as a function of eccentricity. Perception **18**, 347–361 (1989)
92. Snedecor, G., Cochran, W.: Statistical methods, 8-th edn. Iowa State University Press, Ames, IA (1989)
93. Solway, A., Botvinick, M.M.: Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. Psychological review **119**(1), 120 (2012)
94. Sprague, N., Ballard, D.: Eye movements for reward maximization. In: Advances in neural information processing systems, vol. 16. MIT Press, Cambridge, MA (2003)
95. Sprenger, A., Friedrich, M., Nagel, M., Schmidt, C.S., Moritz, S., Lencer, R.: Advanced analysis of free visual exploration patterns in schizophrenia. Frontiers in psychology **4** (2013)
96. Stephen, D., Mirman, D., Magnuson, J., Dixon, J.: Lévy-like diffusion in eye movements during spoken-language comprehension. Physical Review E **79**(5), 056,114 (2009)
97. Strasburger, H., Rentschler, I., Jüttner, M.: Peripheral vision and pattern recognition: A review. Journal of Vision **11**(5) (2011)
98. Sun, Y., Fisher, R., Wang, F., Gomes, H.M.: A computer vision model for visual-object-based attention and eye movements. Computer Vision and Image Understanding **112**(2), 126 – 142 (2008)
99. Tatler, B., Baddeley, R., Vincent, B.: The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. Vision research **46**(12), 1857–1862 (2006)
100. Tatler, B., Hayhoe, M., Land, M., Ballard, D.: Eye guidance in natural vision: Reinterpreting salience. Journal of vision **11**(5) (2011)
101. Tatler, B., Vincent, B.: Systematic tendencies in scene viewing. Journal of Eye Movement Research **2**(2), 1–18 (2008)
102. Tatler, B., Vincent, B.: The prominence of behavioural biases in eye guidance. Visual Cognition **17**(6-7), 1029–1054 (2009)
103. Toh, W.L., Rossell, S.L., Castle, D.J.: Current visual scanpath research: a review of investigations into the psychotic, anxiety, and mood disorders. Comprehensive psychiatry **52**(6), 567–579 (2011)
104. Torralba, A.: Contextual priming for object detection. Int. J. of Comp. Vis. **53**, 153–167 (2003)
105. Treisman, A.: Feature binding, attention and object perception. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **353**(1373), 1295–1306 (1998)
106. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive psychology **12**(1), 97–136 (1980)
107. Underwood, G., Foulsham, T.: Visual saliency and semantic incongruency influence eye movements when inspecting pictures. The Quarterly journal of experimental psychology **59**(11), 1931–1949 (2006)
108. Underwood, G., Foulsham, T., van Loon, E., Humphreys, L., Bloyce, J.: Eye movements during scene inspection: A test of the saliency map hypothesis. European Journal of Cognitive Psychology **18**(03), 321–342 (2006)
109. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and Vision Computing **27**(12), 1743–1759 (2009)
110. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision **57**(2), 137–154 (2004)
111. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks **19**(9), 1395–1407 (2006)
112. Wang, H., Pomplun, M.: The attraction of visual attention to texts in real-world scenes. Journal of Vision **12**(6) (2012)
113. Wilming, N., Harst, S., Schmidt, N., König, P.: Saccadic momentum and facilitation of return saccades contribute to an optimal foraging strategy. PLoS Comput. Biol. **9**(1), e1002,871 (2013)
114. Wischnewski, M., Belardinelli, A., Schneider, W., Steil, J.: Where to Look Next? Combining Static and Dynamic Proto-objects in a TVA-based Model of Visual Attention. Cognitive Computation **2**(4), 326–343 (2010)
115. Wolfe, J.M.: Guided search 2.0 a revised model of visual search. Psychonomic bulletin & review **1**(2), 202–238 (1994)
116. Wolfe, J.M.: When is it time to move to the next raspberry bush? foraging rules in human visual search. Journal of Vision **13**(3) (2013). DOI 10.1167/13.3.10. URL `http://www.journalofvision.org/content/13/3/10.abstract`
117. Zelinsky, G.J.: A theory of eye movements during target acquisition. Psychological review **115**(4), 787 (2008)