Ecological Sampling of Gaze Shifts

Giuseppe Boccignone and Mario Ferraro

***Abstract*—Visual attention guides our gaze to relevant parts of the viewed scene, yet the moment-to-moment relocation of gaze can be different among observers even though the same locations are taken into account. Surprisingly, the variability of eye movements has been so far overlooked by the great majority of computational models of visual attention.**

In this paper we present the Ecological Sampling model, a stochastic model of eye guidance explaining such variability. The gaze shift mechanism is conceived as an active random sampling that the "foraging eye" carries out upon the visual landscape, under the constraints set by the observable features and the global complexity of the landscape. By drawing on results reported in the foraging literature, the actual gaze relocation is eventually driven by a stochastic differential equation whose noise source is sampled from a mixture of α-stable distributions.

This way, the sampling strategy proposed here allows to mimic a fundamental property of the eye guidance mechanism: where we choose to look next at any given moment in time is not completely deterministic, but neither is it completely random

To show that the model yields gaze shift motor behaviors that exhibit statistics similar to those displayed by human observers, we compare simulation outputs with those obtained from eye-tracked subjects while viewing complex dynamic scenes.

***Index Terms*—Visual attention, eye movements, salience, α-stable processes, Lévy flight, foraging.**

I. Introduction

In this paper we shall consider the problem of the variability of visual scanpaths (the sequence of gaze shifts) produced by human observers. When looking at natural movies under a free-viewing or a general-purpose task, the relocation of gaze can be different among observers even though the same locations are taken into account. In practice, there is a small probability that two observers will fixate exactly the same location at exactly the same time. Such variations in individual scanpaths (as regards chosen fixations, spatial scanning order, and fixation duration) still hold when the scene contains semantically rich "objects". Variability is even exhibited by the same subject along different trials on equal stimuli. Further, the consistency in fixation locations between observers decreases with prolonged viewing [1]. This effect is remarkable when free-viewing static images: consistency in fixation locations selected by observers decreases over the course of the first few fixations after stimulus onset [2] and can become idiosyncratic.

Challenges: Although the ability to predict where a human might fixate elements of a viewed scene has long been of interest in the computational vision community [3], [4], the problem in question has hitherto been overlooked. Indeed, a computational model of visual attention and eye guidance should predict where will the eyes select the target of the next fixation by providing: i) a mapping *viewed scene* $\mapsto$ *gaze sequence*; ii) a procedure that implements such mapping. One paradigmatic example is the most prominent model in the literature proposed by Itti *et al* [5]. In this model, attention deployment is explained in terms of visual salience as the output of a competitive process between a set of basic contrast features. Eye guidance is conceived as a Winner-Take-All (WTA) selection of most salient locations.

Nevertheless, most approaches focus on computing a mapping from an image, or, less frequently, from an image sequence to a representation suitable to ground the eye guidance process (e.g., see the recent review by Borji and Itti [4]). Such representation is typically shaped in the form of a saliency map, which is derived either bottom-up, as in [5], or top-down modulated by cognitive and contextual factors (e.g., [6], [7]). The saliency map is then 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 [4]. The problem of eye guidance is somehow neglected or, if needed for practical purposes [8], it is solved by adopting some deterministic choice procedure. The latter is usually based on the $\arg\max$ operation [9]. The aforementioned WTA scheme [5], [9], or the selection of the proto-object with the highest attentional weight [10] are two examples. Even when probabilistic frameworks are used to infer where to look next, the final decision is often taken via the maximum a posteriori (MAP) criterion, which again is an $\arg\max$ operation (e.g., [11]–[15]), or variants such as the robust mean (arithmetic mean with maximum value) over candidate positions [16].

Thus, as a matter of fact, the majority of models that have been proposed so far (with few notable exceptions discussed afterward), hardly take into account one fundamental feature characterizing human oculomotor behavior: where we choose to look next at any given moment in time is not completely deterministic, but neither is it completely random [17]. Indeed, even though the partial mapping *viewed scene* $\mapsto$ *salience* is taken for granted (which could be questioned under some circumstances, [2]), current accounts of the subsequent step, i.e. *salience* $\mapsto$ *gaze sequence*, are still some way from explaining the complexities of eye guidance behavior. In the work presented here we attempt at filling this gap.

Our approach: We assume that the gaze sequence is generated by an underlying stochastic process, accounting for several factors involved in the guidance of eye-movements (e.g., stochastic variability in neuromotor force pulses [18], systematic tendencies in oculomotor behavior [19], see Section II).

The ultimate aim of the present study is to develop a model

G. Boccignone is with the Dipartimento di Informatica, Universitá di Milano, via Comelico 39/41, Milano, Italy
E-mail: (see http://boccignone.di.unimi.it).

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

that describes statistical properties of gaze shifts as closely as possible. Experimental findings have shown that human gaze shift amplitude distributions are positively skewed and long-tailed (e.g., [19]). Drawing on results reported in the foraging literature, where similar distributions characterize the moment-to-moment relocation of many animal species between and within food patches [20], [21], we introduce a composite random walk model for the "foraging eye", which we name *Ecological Sampling* (ES).

The ES scheme, discussed in Section III, models the ecological exploration undertaken by the "foraging eye" while stochastically sampling a complex time-varying visual landscape (here an image sequence) represented in terms of information patches. In the ES model the eye guidance strategy amounts to choose where to look next by sampling the appropriate motor behavior (i.e., the action to be taken: fixating, pursuing or saccading), conditioned on the perceived world and on previous action. More precisely, the appropriate oculomotor behavior is sampled from a mixture of α-stable distributions. The choice and the execution of the oculomotor behavior depend upon both the local information properties of patches and their global configuration within the time-varying landscape (complexity).

To show that the model yields gaze shift motor behaviors that exhibit statistics similar to those pertaining to human observers, in Section IV we compare ES outputs with those obtained from eye-tracked subjects viewing complex videos and collected in a publicly available dataset.

Contributions: The main contributions of this paper lie in the following. 1) A novel and general probabilistic framework for eye guidance on complex time-varying scenes is provided, which revises an early conjecture presented in [22] and grounds its assumptions on empirical analysis of eye-tracked data. 2) The ES guidance mechanism can mimic variability in scanpaths close to that exhibited by human subjects. 3) The scanpath results from the composition of random walks whose stochastic part is driven by different α-stable components. This allows to treat different types of eye movements within the same framework, thus making a step towards the unified modeling of different kinds of gaze shifts, which is a recent trend in eye movement research [23], [24]. 4) The gaze is deployed at patches, i.e. proto-objects, rather than points (differently from [22]). Thus, the eye guidance mechanism could be straightforwardly integrated with a probabilistic object or context-based visual attention scheme [6], [7].

II. Background

Eye movements such as saccades and smooth pursuit, followed by fixations, play an important role in human vision. They allow high-spatial-frequency *sampling* of the visual environment by controlling the direction of the foveal projections (the center of best vision) of the two eyes [23]. Frequent saccades avoid building detailed models of the whole scene [2] and are a characteristic mode of exploratory movements across a wide range of species and types of visual systems.

The pursuit system uses information about the speed of a moving object to produce eye movements of comparable speed, thereby keeping the image of the object on or near the fovea.

Fixations themselves are not simply the maintenance of the visual gaze on a single location but rather a slow oscillation of the eye [23]. They are never perfectly steady and different mechanisms can be at their origin, e.g., microsaccades [25]. Thus eye fixations are better defined as the amount of continuous time spent looking within a circumscribed region (e.g., minimum 50 milliseconds within a spatially limited region, typically $0.5 - 2.0^\circ$ degrees of visual angle [26]).

The variability characterizing *how* we move the eyes occurs ubiquitously, and it may mediate a variety of motor and perceptual phenomena [3], [19]. At a low-level, variability in motor responses originates from endogenous stochastic variations that affect each stage between a sensory event and the motor response [18]. At this level the issue of stochasticity in scanpaths, debated in early studies [27], [28], may be more generally understood on the basis that randomness assumes a fundamental role in adaptive optimal control of gaze shifts; in this perspective, variability is an intrinsic part of the optimal control problem, rather than being simply "noise" [29].

At a higher level it might reflect the individual's learnt knowledge of the structure of the world, the distribution of objects of interest, and task parameters. The latter factors can be summarized in terms of oculomotor tendencies or biases [19]. Systematic tendencies in oculomotor behavior can be thought of as regularities that are common across all instances of and manipulations to the behavior. Under certain conditions these provide a signature of the oculomotor behavior peculiar to an individual (the idiosyncrasy of scanpaths [2], [30]). Oculomotor biases can also be considered as mechanisms tied to strategies that are optimal to minimize search time and maximize accuracy [31].

Tatler and Vincent in their elegant study [19] were the first to show that exploiting these oculomotor biases, the performance of a salience model can be improved from 56% to 80% by including the probability of saccade directions and amplitudes. Strikingly, they found evidence that a model based on oculomotor biases alone performs better than the standard salience model. However, they did not provide neither a formal characterization 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.

Such tendencies can be detected in saccade amplitudes, which show a positively skewed, long-tailed distribution in most experimental settings in which complex scenes are viewed [19]. Similarly, long-tailed distributions have been recently reported on natural movies [1].

More generally, the idea of inferring, through sampling, the properties of a surrounding, uncertain world (either a natural landscape or a fictitious one such as a probability distribution) can be related to the notion of random walk biased by an external force field. In continuous time a d-dimensional random motion of a point, with stochastic position $\mathbf{r}(t)$, under the influence of a force field can be described by

the Langevin stochastic equation [32]

$$d\mathbf{r}(t) = \mathbf{g}(\mathbf{r}, t)dt + \mathbf{D}(\mathbf{r}, t)\boldsymbol{\xi}dt. \quad (1)$$

The trajectory of the variable $\mathbf{r}$ is determined by a deterministic part $\mathbf{g}$, the drift, and a stochastic part $\mathbf{D}(\mathbf{r}, t)\boldsymbol{\xi}dt$, where $\boldsymbol{\xi}$ is a random vector and $\mathbf{D}$ is a weighting factor. Note that in many applications [33] $\mathbf{g}(\mathbf{r}, t)$ is modeled as a force field due to a potential $V(\mathbf{r}, t)$, that is $\mathbf{g}(\mathbf{r}, t) = -\nabla V(\mathbf{r}, t)$.

The stochastic part of the motion is determined by the probability density function f from which $\boldsymbol{\xi}$ is sampled, and different types of motion can be generated by resorting to the class of the so called α-stable distributions [34]. These form a four-parameter family of continuous probability densities, say $f(\boldsymbol{\xi}; \alpha, \beta, \gamma, \delta)$. The parameters are the skewness β (measure of asymmetry), the scale γ (width of the distribution) and the location δ and, most important, the characteristic exponent α, or index of the distribution that specifies the asymptotic behavior of the distribution. The relevance of α derives from the fact that the probability density function (pdf) of jump lengths scales, asymptotically, as $l^{-1-\alpha}$. Thus, relatively long jumps are more likely when α is small. By sampling $\boldsymbol{\xi} \sim f(\boldsymbol{\xi}; \alpha, \beta, \gamma, \delta)$, for $\alpha \geq 2$ the usual random walk (Brownian motion) occurs; if $\alpha < 2$, the distribution of lengths is "broad" and the so called Levy flights take place.

In a seminal paper [35], Brockmann and Geisel argued that a visual system producing Lévy flights implements a more efficient strategy of shifting gaze in a random visual environment than any strategy employing a typical scale in gaze shift magnitudes. Further evidence of Lévy diffusive behavior of scanpaths has been presented in [36]. Potential functions in a Langevin equation have been first used in [33], to address scanpath generation in the framework of a *foraging* metaphor.

Indeed, the heavy-tailed distributions of gaze shift amplitudes are close to those characterizing the foraging behavior of many animal species. Lévy flights have been used to model optimal searches of foraging animals, namely their moment-to-moment relocations/flights used to sample the perceived habitat [20]. However, the general applicability of Lévy flights in ecology and biological sciences is still open to debate. In complex environments, optimal searches are likely to result from a mixed/composite strategy, in which Brownian and Levy motions can be adopted depending on the structure of the landscape in which the organism moves [21]. Lévy flights are best suited for the location of randomly, sparsely distributed patches and Brownian motion gives the best results for the location of densely but random distributed within-patch resources [37].

A preliminary attempt towards a composite sampling strategy for modelling gaze shift mechanisms has been presented in [22]. However, that approach only conjectured a simple binary switch between a Gaussian and a Cauchy-like walk. While providing some promising results, the approach lacked of a general framework and did not ground its assumptions on empirical analysis of eye-tracked data. In the work presented here, experimental data analysis has been exploited to substantially revise [22] and to formulate the general ES model detailed in the following Section.

Notations: The notations used in Section III are listed in the following:

$\mathbf{I}(t)$	a snapshot of the raw time-varying natural habitat at time t, i.e., a frame of the input video $\mathbf{I}$;
$\mathbf{F}(t)$	the observable features of the habitat;
$\mathcal{W}(t)$	the set of random variables (RV) characterizing the perceived time-varying natural habitat;
$\mathcal{A}(t)$	the set of RVs characterizing an oculomotor behavior, briefly, the action within the habitat;
$\mathcal{S}(t)$	the set of RVs characterizing the salience landscape of the habitat;
$\mathcal{O}(t)$	the set of RVs characterizing the patches of the habitat;
$\mathcal{M}(t)$	the patch map
L	the spatial support of the video frame $\mathbf{I}(t)$;
$\mathbf{r}(t)$	a point of coordinates $(x, y) \in L$;
$\mathbf{r}_F(t)$	the gaze fixation position at time t, i.e. the Focus of Attention (FOA) center;
$s(\mathbf{r}, t)$	a binary r. v. labelling location $\mathbf{r} \in L$ as salient or non salient;
N_P	total number of patches;
θ_p	shape parameters of patch p, i.e., location μ_p and covariance Σ_p ;
$m_p(\mathbf{r}, t)$	a binary RV labelling location $\mathbf{r} \in L$ as belonging or not to patch p;
$N_{i,p}$	total number of interest points generated from patch p;
$\mathbf{r}_{i,p}$	the i-th interest point generated from patch p;
$x(1:t)$	shorthand notation for the temporal sequence $x(1), x(2), \cdots, x(t)$;
K	the number of possible actions;
k	action index, in the range $[1, \cdots, K]$;
$z(t)$	categorical RV taking values in $[1, \cdots, K]$;
$\pi_k(t)$	probability of choosing action k at time t;
$\pi(t)$	the set of probabilities $\{\pi_k(t)\}_{k=1}^K$;
$\nu_k(t)$	hyper-parameter of the Dirichlet distribution over $\pi_k(t)$;
$\nu(t)$	the set of hyperparameters $\{\nu_k(t)\}_{k=1}^K$;
$w(\mathbf{r}_c)$	a cell or window, centered at $\mathbf{r}_c$, i.e., the elementary unit to partition the support L in the configuration space;
N_w	the number of cells in the configuration space;
$H(t)$	the Boltzmann-Gibbs-Shannon entropy of the configuration space;
$\Omega(t)$	the order parameter;
$\Delta(t)$	the disorder parameter;
$\mathcal{C}(t)$	the complexity index;
η_k	the set of parameters $\alpha_k, \beta_k, \gamma_k, \delta_k$ shaping the α-stable distribution tied to action k;
$\boldsymbol{\xi}_k$	random vector of components $\xi_{k,j}$ sampled from the k-th α-stable distribution;
N_V	the number of gaze attractors.

III. The Ecological Sampling model

Let us assume that, at time t, the gaze position is set at $\mathbf{r}_F(t)$ (the center of the focus of attention, FOA). The ES strategy is part of the action/perception cycle undertaken by the observer and amounts to choose where to look next, i.e. $\mathbf{r}_F(t+1)$, by sampling the appropriate motor behavior, or action $\mathcal{A}(t)$, conditioned on the perceived world $\mathcal{W}(t)$ and on previous action $\mathcal{A}(t-1)$. At the most general level it can be articulated in the following steps:

1) Sampling the natural habitat:

$$\mathcal{W}^*(t) \sim P(\mathcal{W}(t)|\mathbf{r}_F(t), \mathbf{F}(t), \mathbf{I}(t)); \quad (2)$$

2) Sampling the appropriate motor behavior:

$$\mathcal{A}(t)^* \sim P(\mathcal{A}(t)|\mathcal{A}(t-1), \mathcal{W}^*(t)); \quad (3)$$

3) Sampling where to look next:

$$\mathbf{r}_F(t+1) \sim P(\mathbf{r}_F(t+1)|\mathcal{A}(t)^*, \mathcal{W}^*(t), \mathbf{r}_F(t)). \quad (4)$$

Here, $P(\mathcal{W}(t)|\mathbf{r}_F(t), \mathbf{F}(t), \mathbf{I}(t))$ represents the world likelihood as gauged through features $\mathbf{F}(t)$ derived from the physical stimulus $\mathbf{I}(t)$, which in turn is foveated at location $\mathbf{r}_F(t)$; $P(\mathcal{A}(t)|\mathcal{W}(t), \mathbf{r}_F(t))$ is the probability of undertaking action $\mathcal{A}(t)$ given the current state of affairs $\mathcal{W}(t)$, and previous behavior $\mathcal{A}(t-1)$. Finally, $P(\mathbf{r}_F(t+1)|\mathcal{A}(t), \mathcal{W}(t), \mathbf{r}_F(t))$ accounts for the gaze shift dynamics, that is the probability of the transition $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t+1)$.

A. Sampling the natural habitat

The "foraging" eye, by gazing at $\mathbf{r}_F(t)$, allows the observer to gauge, at time t, the physical world through features $\mathbf{F}(t)$. Differently from [22], the visible features serve the purpose of structuring the habitat $\mathcal{W}(t)$ in terms of a *landscape* $\mathcal{S}(t)$ and a set of landscape *patches* $\mathcal{O}(t)$, i.e. $\mathcal{W}(t) = \{\mathcal{S}(t), \mathcal{O}(t)\}$.

The landscape is defined as a map of spatially interesting/uninteresting locations $\mathcal{S}(t) = \{s(\mathbf{r}, t)\}_{\mathbf{r}\in L}$. Following [38], we use $s(\mathbf{r}, t)$ as a binary random variable (RV) to label point $\mathbf{r}$ as salient or non salient.

Under this assumption, the posterior $P(\mathcal{W}(t)|\mathbf{r}_F(t), \mathbf{F}(t), \mathbf{I}(t))$ in (2) can be factorized as $P(\mathcal{O}(t), \mathcal{S}(t)|\mathbf{r}_F(t), \mathbf{F}(t), \mathbf{I}(t)) = P(\mathcal{O}(t)|\mathcal{S}(t))P(\mathcal{S}(t)|\mathbf{r}_F(t), \mathbf{F}(t), \mathbf{I}(t))$.

The probability $P(\mathcal{S}(t)|\mathbf{F}(t), \mathbf{I}(t), \mathbf{r}_F(t))$ represents the saliency map of such landscape, evaluated under the feature matrix $\mathbf{F}(t)$, which is in turn obtained from $\mathbf{I}(t)$ gazed at $\mathbf{r}_F(t)$ and thus foveated at that position. The foveated frame $\widehat{\mathbf{I}}$ is calculated by blurring the current frame using a Gaussian function centered at $\mathbf{r}_F(t)$. Eventually, the feature matrix is obtained $\mathbf{F} = \mathbf{F}(\widehat{\mathbf{I}})$.

Such definition of saliency as a posterior probability on locations is common to many methods in the literature (e.g., see [38] for bottom-up saliency computation or [6], for a general top-down, object-based method). It is worth noting that the model presented here needs not to rely on any specific method for computing saliency.

Patches may be conceived in terms of foraging sites around which food items (or moving preys) can be situated [39]. In the visual attention realm, patches can stand for generic *proto-objects* [9], [10], [38], [40], [41].

Thus, at any given time t, the observer perceives a set $\mathcal{O}(t)$ of a number patches in terms of prey clusters, each patch being characterized by different shape and location. More formally, $\mathcal{O}(t) = (O(t), \Theta(t))$, where $O(t) = \{O_p(t)\}_{p=1}^{N_P}$ is the ensemble of patches and $\Theta(t)$ their parametric description.

In particular, $O_p(t) = \{\mathbf{r}_{i,p}\}_{i=1}^{N_{i,p}}$ is a sparse representation of patch p as the cluster of interest points (preys, food items) that can be sampled from it. Patch sampling is driven by the locations and the shapes of the habitat patches described through the set of parameters $\Theta(t) = \{\Theta_p(t)\}_{p=1}^{N_P}$.

More precisely, each patch is parametrized as $\Theta_p(t) = (\mathcal{M}_p(t), \theta_p)$. 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 patches within the habitat at time t is given by $\mathcal{M}(t) = \bigcup_{p=1}^{N_P} \mathcal{M}_p(t)$. This map may be derived either by simple segmentation techniques of the saliency map [38], [9], [41], or by exploiting higher level cues [6].

The patch map provides the necessary spatial support for a 2D ellipse approximation of each patch, whose location and shape are parametrized as $\theta_p = (\mu_p, \Sigma_p)$ [10].

This way, the term $P(\mathcal{O}(t)|\mathcal{S}(t))$ can be factorized as $P(O(t), \theta(t), \mathcal{M}(t)|\mathcal{S}(t)) = P(O(t)|\theta(t), \mathcal{M}(t), \mathcal{S}(t))\ P(\theta(t)|\mathcal{M}(t), \mathcal{S}(t))\ P(\mathcal{M}(t)|\mathcal{S}(t))$.

Eventually, by assuming independent patches, the first sampling step (2) boils down to the following sub-steps:

$$\mathcal{S}^*(t) \sim P(\mathcal{S}(t)|\mathbf{F}(\widehat{\mathbf{I}(t)})); \quad (5)$$

$$\mathcal{M}^*(t) \sim P(\mathcal{M}(t)|\mathcal{S}^*(t)); \quad (6)$$

for $p = 1, \cdots, N_P$

$$\theta_p^*(t) \sim P(\theta_p(t)|\mathcal{M}_p^*(t) = 1, \mathcal{S}^*(t)), \quad (7)$$

$$O_p^*(t) \sim P(O_p(t)|\theta_p^*(t), \mathcal{M}_p^*(t) = 1, \mathcal{S}^*(t)). \quad (8)$$

The first sub-step samples the foveated salience map. The second samples the patch map from the landscape. The third derives patch parameters $\theta(t)_p = (\mu_p(t), \Sigma_p(t)))$.

Eventually, sub-step (8) generates clusters of interest points on the landscape, one cluster for each patch. By assuming a Gaussian distribution centered on the patch, i.e. $P(\mathbf{r}_p|\theta_p(t), \mathcal{M}_p(t), \mathcal{S}(t)) = \mathcal{N}(\mathbf{r}_p; \mu_p(t), \Sigma_p(t))$, Eq. (8) can be further specified as:

$$\mathbf{r}_{i,p} \sim \mathcal{N}(\mathbf{r}_p; \mu_p(t), \Sigma_p(t)), i = 1, \cdots, N_{i,p}. \quad (9)$$

Thus, the set of all interest points characterizing the habitat can be obtained as $O(t) = \bigcup_{p=1}^{N_P} \{\mathbf{r}_{i,p}(t)\}_{i=1}^{N_{i,p}}$. Note that $O(t)$ provides a sparse representation of the original saliency map, since $|O(t)| = N_s = N_{i,p} \times N_p \ll |L|$.

B. Sampling the appropriate motor behavior

We represent the process of selecting the most appropriate motor behavior, which we briefly call an *action*, as a two-component process unfolding in time: the actual selection and the evolution of parameters governing such selection. More formally, an action is the pair $\mathcal{A}(t) = (z(t), \pi_t)$, where $z(t)$

is a categorical RV with K states $z(t) = \{z(t) = k\}_{k=1}^{K}$, each state being one possible action. The probabilities of choosing one of K behaviors $\pi(t) = \{\pi_k(t)\}_{k=1}^{K}$ are the parameters governing the multinomial choice of $z(t)$.

By letting the action choice $\mathcal{A}(t)$ depend only on the sampled interest points, then, we can factorize $P(\mathcal{A}(t)|\mathcal{A}(t-1), O(t)) = P(z(t), \pi(t)|z(t-1), \pi(t-1), O(t)) = P(z(t)|\pi(t))P(\pi(t)|\pi(t-1), O(t))$.

Since in our case, differently from [22], the motor behavior is chosen among K possible kinds, $P(z|\pi)$ is the Multinomial distribution $Mult(z(t)|\pi(t)) = \prod_{k=1}^{K} [\pi_k(t)]^{z_k(t)}$ with $\pi_k = P(z = k|\pi)$.

The conjugate prior of the latter is the Dirichlet distribution, $P(\pi(t)) = Dir(\pi(t); \nu(t)) = \frac{\Gamma(\sum_k \nu_k(t))}{\prod_k \Gamma(\nu_k(t))} \prod_k \pi_k(t)^{\nu_k(t)-1}$, where $\Gamma(\cdot)$ is the Gamma function.,

Note that the transition $\mathcal{A}(t-1) \rightarrow \mathcal{A}(t)$, is governed by the posterior transition density $P(\pi(t)|\pi(t-1), O(t))$. Since here we are dealing with a kind of (discrete time) dynamical system, this represents the transition over a time slice, that is an instance of the process that actually has been running up to time t.

Under first-order Markov assumption [42], the posterior pdf can be fully written as $P(\pi(t)|\pi(t-1), O(1:t)) \propto P(O(t)|\pi(t))P(\pi(t-1)|O(1:t-1))$. Such recursive updating can be analytically specified, in the case of the Dirichlet distribution, by the hyper-parameter update

$$\nu_k(t) = \nu_k(0) + N_k(t), \tag{10}$$

where, in Iverson's notation, $N_k(t) = N(t)\,[E = k]$ is a count on events depending on the sparse representation $O(t)$. To make this statement explicit, we will write $P(\pi(t)|\nu(t), O(t)) = P(\pi(t)|\nu(O(t)))$ to remark the dependance of the hyperparameters on $O(t)$.

Instead of using the configuration of $O(t)$ as the explanatory variable influencing the motor behavior choice, we will use a dependent variable, a global parameter, say $\mathcal{C}(O(t))$, providing at a glance the "gist" of the spatio-temporal configuration of the landscape. One such outcome variable is the spatio-temporal heterogeneity of the landscape.

For instance, in ecological modelling [43] a widely adopted measure to gauge the heterogeneity is the landscape entropy determined by dispersion/concentration of food items or preys. Here, generalizing this approach, we use $\mathcal{C}(O(t))$ (or more simply $\mathcal{C}(t)$) to capture the time-varying configurational complexity of interest points within the landscape.

Following Shiner *et al.* [44], the complexity $\mathcal{C}(t)$ can be defined in terms of order/disorder of the system:

$$\mathcal{C}(t) = \Delta(t) \cdot \Omega(t), \tag{11}$$

where $\Delta \equiv H/H_{sup}$ is the disorder parameter, $\Omega = 1 - \Delta$ is the order parameter, and H the Boltzmann-Gibbs-Shannon (BGS) entropy of the system with H_{sup} its supremum.

Eq. (11) embodies the general principle underlying all approaches undertaken to define the complexity of a dynamic system: complex systems are neither completely random neither perfectly ordered and complexity should reach its maximum at a level of of randomness away from these extremes.

In the case of a time-varying visual landscape, a crowded scene with many people moving represents a disordered system (high entropy, low order) as opposed to a static scene where no events take place (low entropy, high order). The highest complexity is reached when specific events occur: two persons meeting at a cross-road while a cyclist is passing by, etc. What is observed in eye-tracking experiments on videos [1] is that low complexity scenarios usually lead to longer flights (saccadic behavior) so as to promptly gather more information, whilst at the edge of order/disorder more complex and mixed behaviors take place (e.g., intertwining fixations, smooth-pursuit, and saccades). To formalize the relationship between the complexity of the habitat and the choice of behavior we proceed as follows.

We compute the BGS entropy H as a function of the spatial configuration of the sampled interest points. The spatial domain L is partitioned into a configuration space of cells (rectangular windows), i.e., $\{w(\mathbf{r}_c)\}_{c=1}^{N_w}$, each cell being centered at $\mathbf{r}_c$. By assigning each interest point to the corresponding window, the probability for point $\mathbf{r}_s$ to be within cell c at time t can be estimated as $P(c,t) \simeq \frac{1}{N_s} \sum_{s=1}^{N_s} \chi_{s,c}$, where $\chi_{s,c} = 1$ if $\mathbf{r}_s \in w(\mathbf{r}_c)$ and 0 otherwise (see, Section IV, for further details).

Thus, $H(t) = -k_B \sum_{c=1}^{N_w} P(c,t) \log P(c,t)$, and (11) can be easily computed. Since we are dealing with a fictitious thermodynamical system, we set Boltzmann's constant $k_B = 1$. The supremum of $H(t)$ is obviously $H_{sup} = \ln N_w$ and it is associated to a completely unconstrained process, that is a process where $H(t) = const$, since with reflecting boundary conditions the asymptotic distribution is uniform.

Given $\mathcal{C}(t)$, we partition the complexity range in order to define K possible complexity events $\{E_{\mathcal{C}(t)} = k\}_{k=1}^{K}$. This way the hyper-parameter update (10) can be rewritten as the recursion

$$\nu_k(t) = \nu_k(t-1) + \left[E_{\mathcal{C}(t)} = k\right], k = 1, \cdots, K. \tag{12}$$

As previously discussed, three possible events will be eventually identified (see Section IV) to provide the gist of the spatio-temporal habitat: "ordered dynamics", "edge dynamics" and "disordered dynamics", each biasing the process toward a specific gaze shift behavior as observed in eye-tracked data [1].

Summing up, the action sampling step (3) amounts to: i) computing the complexity of the landscape as a function of sampled interest points $O(t)$; ii) updating accordingly the hyperparameters $\nu_k(O(t))$ (12); iii) sampling the action $\mathcal{A}^*(t)$ as:

$$\pi^*(t) \sim Dir(\pi|\nu(O(t))); \tag{13}$$

$$z^*(t) \sim Mult(z(t)|\pi^*(t)). \tag{14}$$

C. Sampling where to look next

Given action $\mathcal{A}^*(t)$, we can rewrite the last sampling step in (4) as:

$$\mathbf{r}_F(t+1) \sim P(\mathbf{r}_F(t+1)|z^*(t) = k, \theta^*(t), \eta, \mathbf{r}_F(t)). \tag{15}$$

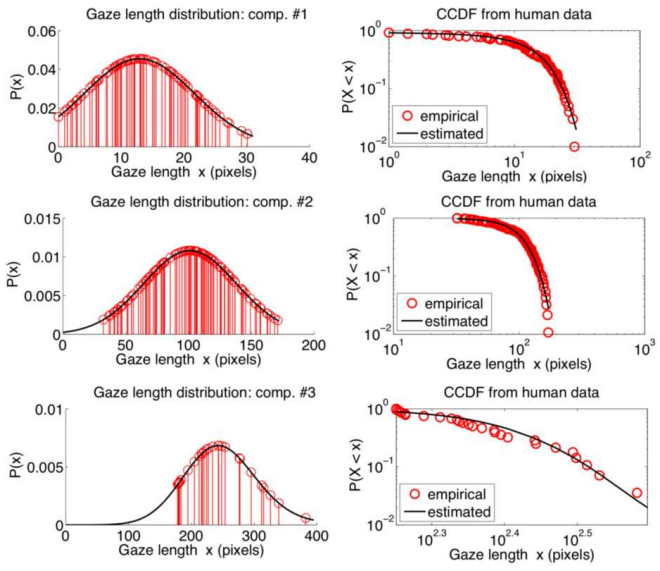

Fig. 1. Results of the α-stable fit of the smooth pursuit and saccadic components for the mtvclip04. The left column figures show the empirical distribution with superimposed the fitted α-stable distributions; the right column figures show the double log-plot of the corresponding CCDF. The top row represents the fitting results for the smooth pursuit component ($\alpha = 2$, $\beta = 1$, $\gamma = 6.20$, $\delta = 12.88$; K-S statistics 0.1200, $p = 0.4431$). The middle row presents the results obtained for the α-stable fit of the first saccadic component ($\alpha = 2$, $\beta = 1$, $\gamma = 26.10$, $\delta = 101.13$; K-S statistics 0.1398, $p = 0.301$). The bottom row presents the results obtained for the second saccadic component ($\alpha = 1.72$, $\beta = 1$, $\gamma = 41.25$, $\delta = 251.25$; K-S statistics 0.1786, $p = 0.7198s$).

Here η play the role of the actual "motor" parameters governing the shift of gaze.

Clearly, the choice among the different oculomotor behaviors follows a Multinomial distribution, $P(\mathbf{r}_F(t+1)|z(t), \theta(t), \eta, \mathbf{r}_F(t)) = \prod_{z(t)} [P(\mathbf{r}_F(t+1)|\mathbf{r}_F(t), \eta)]^{z(t)}$ where $P(\mathbf{r}_F(t+1)|z(t) = k, \theta^*(t), \eta, \mathbf{r}_F(t)) = P(\mathbf{r}_F(t+1)|\theta^*(t), \eta_k, \mathbf{r}_F(t))$ is the oculomotor state transition probability of the shift $\mathbf{r}_F(t) \rightarrow \mathbf{r}_F(t+1)$, which is generated according to motor behavior $z^*(t) = k$ and thus regulated by parameters η_k.

We sample $\mathbf{r}_F(t+1)$ by making explicit the stochastic dynamics behind the process [45]. To this end, Eq. (1) is reformulated as a two-dimensional dynamical system in which the drift term depends on a potential V and the stochastic part is driven by one-of-K possible types of α-stable motion

$$d\mathbf{r}_F(t) = -\nabla V(\mathbf{r}_F, t)dt + \mathbf{D}(\mathbf{r}_F, t)\boldsymbol{\xi}_k(t)dt. \qquad (16)$$

The drift term, the first term on the r.h.s. of (16), is modeled as follows. In a foraging framework, animals are expected to be attracted or repelled from certain sites; therefore $V(\mathbf{r}_F, t)$ can be assumed to depend on the distance between the position $\mathbf{r}_F$ of the animal and the position $\mathbf{r}^*$ of the nearest of such sites. For simplicity, we define $V(\mathbf{r}_F, t) = \frac{1}{2}|\mathbf{r}_F(t) - \mathbf{r}^*(t)|^2$.

Then, we select N_V sites (according to some rule, e.g, the top-N_V most attractive). By assuming that such *attractors* act as independent sources, the gradient of the potential can be eventually obtained from the linear combination of N_V local potentials,

$$-\nabla V(\mathbf{r}_F, t) = -\sum_{p=1}^{N_V}(\mathbf{r}_F(t) - \mathbf{r}_p(t)). \qquad (17)$$

The selection of attractors $\mathbf{r}_p(t)$ clearly depends on the action state k. If a fixation / pursuit behavior has been sampled, these will be chosen as the N_V most valuable points sampled from the current patch, that is $N_V \leq N_{i,p}$. Otherwise, the attractors can be straightforwardly identified with patch centers $\mu_p(t)$, i.e., $N_V = N_P$. The latter are to be considered the possible targets for medium or large shifts of gaze (saccades).

Following [32], the components $\xi_{k,j}$, $j = 1, 2$ are sampled from an α-stable distribution $f(\boldsymbol{\xi}; \eta_k)$ and they are assumed to be statistically independent, so that $\mathbf{D}(\mathbf{r}_F, t)$ is a diagonal matrix. The elements of $\mathbf{D}(\mathbf{r}_F, t)$ can be determined on the basis of theoretical consideration or by the experimental data [32]. Here we have chosen to set the elements of $\mathbf{D}$ equal to the width γ_k of the α-stable distribution characterizing the random walk at time t, namely $\mathbf{D}(\mathbf{r}_F, t) = \gamma_k \mathbb{I}$ with $\mathbb{I}$ the 2×2 identity matrix.

By using these assumptions and by resorting to the Euler-Maruyama discretization [46], for a small time step $\tau = t_{n+1} - t_n$, the SDE (16) is integrated as:

$$\mathbf{r}_F(t_{n+1}) \approx \mathbf{r}_F(t_n) - \sum_{p=1}^{N_V}(\mathbf{r}_F(t_n) - \mathbf{r}_p(t_n))\tau + \gamma_k \mathbb{I}\tau^{1/\alpha_k}\boldsymbol{\xi}_k. \qquad (18)$$

This step provides the explicit procedure for sampling the next gaze shift.

IV. Simulation

Simulations have been carried out to generate statistics of gaze shift behavior of the model. The latter have been compared with those exhibited by human observers (subsection IV-E).

The rationale is that if observed gaze shifts are generated by an underlying stochastic process the distribution functions and the temporal dynamics of eye movements should be completely specified by the stochastic process [47]. At the same time, different stochastic processes often yield different marginal distribution functions in the outcome variables; thus, knowing the precise distribution functions of a RV should suggest plausible generative mechanisms and rule out improbable ones.

Following previous work in the literature [35], the experiments were specifically designed to confront gaze shift magnitude distribution of subjects scanning videos (collected in a publicly available dataset, subsection IV-A), with those obtained by running an implementation of the ES model (detailed in subsection IV-C). Indeed, the study of shift amplitude distribution, and in particular of the corresponding complementary cumulative distribution function (CCDF), is the standard convention in the literature of different fields dealing with anomalous random walks such as foraging [21], human mobility [48], statistical physics [49]. In this respect, a preliminary, non trivial problem to solve is to derive from

Fig. 2. The Ecological Sampling implementation at a glance. From top to bottom, left to right: the original frame; the foveated frame; the raw saliency map; detected patches; sampled interest points (drawn as white disks for visualization purpose); the sampled FOA

recorded eye-tracked data the number K of motor behaviors and to infer the related α-stable distribution parameters; to such end a fitting procedure has been devised, which is presented in subsection IV-B.

A. Dataset

We used the CRCNS eye-1 dataset created by University of South California. The dataset is freely available and consists of a body of 520 human eye-tracking data traces recorded (240 Hz sampling rate) while normal, young adult human volunteers watched complex video stimuli (TV programs, outdoors videos, video games), under the generic task of "following main actors and actions". It comprises eye movement recordings from eight distinct subjects watching 50 different video clips (MPEG-1, 640×480 pixels, 30 fps, approximately 25 minutes of total playtime; the Original dataset), and from another eight subjects watching the same set of video clips after scrambling them into randomly re-ordered sets of $1-3$s clippets (the MTV-style dataset). See [50] for a description and https://crcns.org/files/data/eye-1/crcns-eye1-summary.pdf for more details.

B. Gaze shifts statistics

We studied the distributions of gaze magnitudes by analyzing eye-tracking results collected in the CRCNS database To this end, gaze shift samples from all the traces of the same video, regardless of the observers, are aggregated together and used in the same distribution. The assumption is that every observer on the same video has the same statistical "mobility tendency" in terms of gaze shifts; then this aggregation is reasonable because every trace obtained from the same video is subject to the same or similar saliency constraints (i.e. visual landscape). The same technique is used in other studies of Levy walks (e.g., [48]) but also in eye-tracking experiments [2]. In the CRCNS database, eye-tracker samples are individually labelled as fixation, saccade or smooth pursuit, from which it is possible to collect empirical gaze magnitude distributions of eye-tracked subjects. Saccade lengths are straightforward to compute as the Euclidean distance between saccade start/end coordinates. For what concerns smooth pursuit, which indeed represents a kind of Continuous Time Random Walk, since movies were displayed in the original experiment at a rate of 33.185 ms/frame, to be consistent, we subsampled by 8 each smooth pursuit sub-tracks in order to work at a frame-rate basis, thus making feasible to compare with the simulation. The same was done for fixational movements, which have been aggregated with pursuit samples.

Given the empirical distributions of smooth pursuit and saccades, it is possible to individually fit such distributions in order to derive the parameters of the underlying alpha-stable distribution. The quality of the fit is 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. For a more precise description of the tail behavior, 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.

Fig. 1 shows one example of the typical behavior of pursuit and saccade gaze shifts in terms of both the shift magnitude distribution and its corresponding upper tail behavior.

We experimentally found that any attempt to fit a unique α-stable function to the empirical distribution of saccades fails to pass the K-S test. This could be expected by visual inspection of the saccade amplitude histogram, which suggest a mixture of two saccadic behaviors. In order to separate the two processes so to use them in the gaze shift generative process (18), one could resort to an α-stable mixture fitting method. Unfortunately, most of the α-stable mixture treatments that have been developed are either tailored for specific cases (e.g., symmetric distributions, Normal-Cauchy distributions, etc) and often rely on heavy Monte Carlo simulations [51]. Thus, we opted for an indirect but effective technique.

First, we hard-clustered the gaze shift samples into an optimal number of α-stable mixture components via a Variational Bayes Student-t Mixture Model (VBSTMM, see [52] for detailed presentation). The reason for using the t-distribution for identifying components stems from the fact that this distribution might be regarded as the strongest competitor to the α-stable distribution. While the α-stable distribution implies extremely slowly decreasing tails, the t distribution exhibits power tails but has the advantage of finite moments. In a second step, each mixture component was separately

used for α-stable parameter estimation. The estimation of the α-stable distribution is complicated by the aforementioned nonexistence of a closed form pdf. Here we have used the approximated parameter estimator proposed in [53].

As a result, what can be observed is that the component accounting for smooth pursuit and fixations (comp. #1) is readily separated from those explaining saccades; in turn, saccade distribution optimally splits in two α-stable components, a first one, in most cases Gaussian-like $\alpha \approx 2$ (comp. #2) related to saccades of medium length, and a second one (comp. #3) related to saccades of higher magnitude. An example of such pattern is shown in Fig. 1. Interestingly enough, such multi-component statistics for saccades provides a rather different result from those usually reported in the literature when considering static images [35], [33] or conjectured for video analysis [22].

C. Implementation details

In order to implement the first sampling step specified in (5), the saliency map $P(\mathcal{S}(t)|\mathbf{F}(t),\mathbf{I}(t),\mathbf{r}_F(t))$ is derived as follows. Given a fixation point $\mathbf{r}_F(t)$ at time t (the frame center is chosen for $t = 1$), we simulate the foveation process by blurring the current RGB frame $\mathbf{I}(t)$ of the input sequence through a Gaussian function centered at $\mathbf{r}_F(t)$. The foveated frame is obtained as $\widehat{\mathbf{I}}(\mathbf{r},t) = I(\mathbf{r},t)\exp\{-(\mathbf{r}(t) - \mathbf{r}_F(t))\Sigma_{FOA}^{-1}(\mathbf{r}(t) - \mathbf{r}_F(t))^T\}$, where $\Sigma_{FOA} = \sigma^2\mathbb{I}$, $\sigma = |FOA|$. Here $|FOA|$ indicates approximately the radius of a FOA, where $|FOA| \approx 1/8\min[width, height]$ of the frame spatial support L.

The foveated frame $\widehat{\mathbf{I}}(\cdot,t)$, is used to compute feature matrix $\mathbf{F}(t)$ and saliency $P(\mathcal{S}(t)|\mathbf{F}(\widehat{\mathbf{I}}(t)))$ through the Self-resemblance method described in [38]. We initially experimented with the Itti *et al.* [5], the Bayesian Surprise [54] and the Graph-Based Visual Saliency [55] methods. However, Self-resemblance provides comparable performance and meanwhile it can handle both static and space-time saliency detection; it avoids explicit motion estimation and meanwhile is able to cope with camera motion.

Next we approximate the sampling steps (6) and (7) to obtain $\mathcal{M}(t)$ and $\theta_p(t)$ as follows.

The proto-object map $\mathcal{M}(t)$ is simply drawn from $P(\mathcal{S}(t)|\mathbf{F}(\widehat{\mathbf{I}}(t)))$ by deriving a preliminary binary map $\widetilde{\mathcal{M}}(t) = \widehat{m}(\mathbf{r},t)\}_{\mathbf{r}\in L}$, such that $\widehat{m}(\mathbf{r},t) = 1$ if $P(s(\mathbf{r},t)|\mathbf{F}(\widehat{\mathbf{I}}(t))) > T_M$, and $\widehat{m}(\mathbf{r},t) = 0$ otherwise. The threshold T_M is an adaptive threshold similar to the methods proposed in [41] and [38], which is determined as three times the mean saliency $E[\mathcal{S}(t)]$ of the frame [41]. The technique of setting T_M so as to achieve 95% significance level in deciding whether the given saliency values are in the extreme tails of the pdf provides comparable results [38]. Indeed, both procedures are based on the assumption that a salient proto-object is a relatively rare region and thus result in values which are in the tails of $P(\mathcal{S}(t)|\mathbf{F}(\widehat{\mathbf{I}}(t)))$.

Following [9], $\mathcal{M}(t) = \{\mathcal{M}_p(t)\}_{p=1}^{N_P}$ 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 $N_P = 8$ to retain the most important patches.

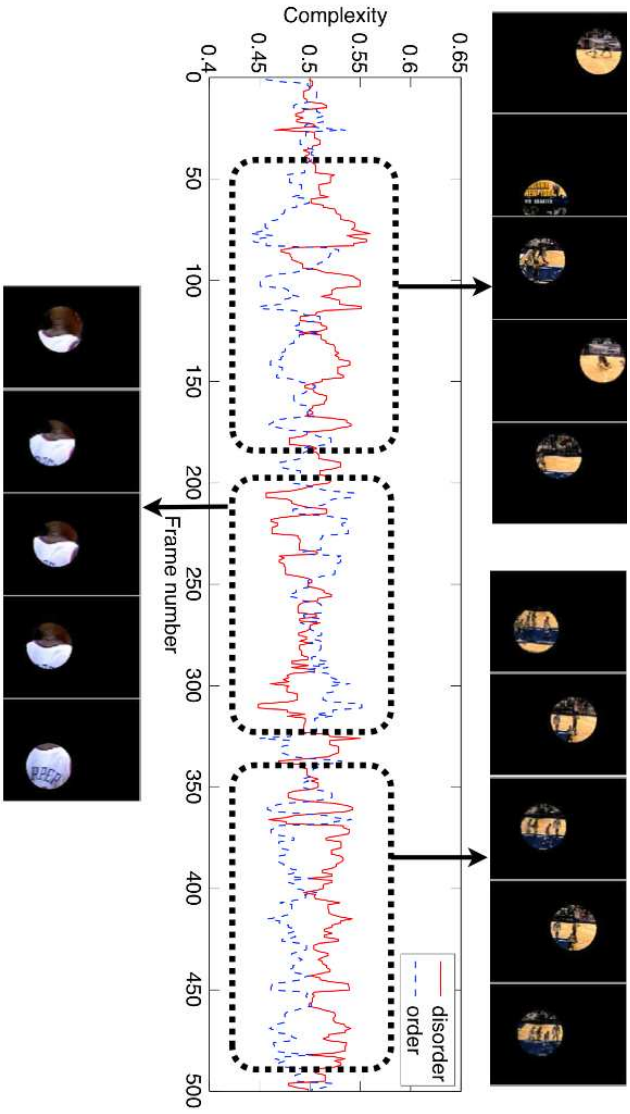

Fig. 3. An example of typical results obtained along the simulation. In the center of the figure the plot shows the evolution of order (dashed line) and disorder parameters Ω and Δ as a function of frame number. From top to bottom, the first dashed box represent a time window where $\Delta > \Omega$ and an excerpt of the resulting saccadic exploratory behavior is shown in the FOA sequence sampling the basket ball actions (top right frame sequence); the second time window reports a switch to a smooth-pursuit regime due to $\Omega > \Delta$ with corresponding foveations on the most important object in the scene (player close-up) shown in the left frame sequence. The successive time window witnesses a new behavioral switch ($\Delta > \Omega$) to a prevalent saccadic explorations of the sport game dynamics (bottom right sequence).

The sampling of patch parameters $\theta_p(t)$ is approximated as follows. By assuming a uniform prior $P(\theta_p(t))$, then $P(\theta_p(t)|\mathcal{M}_p(t),\mathcal{S}(t)) \propto P(\mathcal{M}_p(t),\mathcal{S}(t)|\theta_p(t))$, so that $\theta_p(t)$ reduce to parameters (rather than RVs) that can be estimated via any maximum-likelihood technique. In the simulation this was obtained by adopting the technique by Halır and Flusser [56], because of its numerical stability and computational efficiency (due to non-iterativity). Once parameters $\theta_p(t)$ have been computed, each patch is used to generate interest points in a number proportional to the area of the ellipse describing the patch. 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 (9). 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.

At this point we compute the order/disorder parameters. We use $N_w = 16$ rectangular windows (approximately covering half of the area covered by a FOA), their size depending on the frame size $|L|$. This choice also provides the best trade-off between coarse to fine properties of the configuration space and the number N_s of sampled interest points. The spatial histogram of interest points is used to estimate empirically the cell probability; the latter is then used to calculate the BGS entropy $H(t)$ of the interest point configuration space, and eventually the disorder and order parameters, $\Delta(t)$ and $\Omega(t)$ to be used in Eq. (11) [44].

Note that $\max \mathcal{C}(t)$ is achieved for $\Delta(t) = \Omega(t) = 0.5$, thus $\max \mathcal{C}(t) = 0.25$. By taking into account the results obtained from eye-tracking data analysis, three complexity events $E_{\mathcal{C}} \in \{1, 2, 3\}$ are devised, which characterize corresponding motor behaviors $k \in \{1, 2, 3\}$: $E_{\mathcal{C}} = 1$ if $\Omega(t) > \Delta(t)$ and $\mathcal{C} < \max \mathcal{C} - \epsilon$ indicating an "ordered dynamics" of the spatio-temporal habitat; $E_{\mathcal{C}} = 3$ if $\Omega(t) < \Delta(t)$ and $\mathcal{C} < \max \mathcal{C} - \epsilon$ for "disordered dynamics"; event $E_{\mathcal{C}} = 2$ occurs within higher range of complexity, $|\mathcal{C} - \max \mathcal{C}| \leqslant \epsilon$ where "edge dynamics" will take place. In the simulation the range value $\epsilon = 0.01$ has been experimentally determined. The empirical consequence of such event detection procedure is that an ordered dynamics of the habitat will most likely bias the shift dynamics toward quasi-Brownian shifts (fixation / pursuit regime), whilst in highly disordered environment, longer shifts are more likely to occur (saccadic regime); at the edge between these regimes, where complexity is high since order is dynamically competing with disorder, $\Omega(t) \approx \Delta(t)$, intermediate length shifts and mixed behaviors will take place

TABLE I
GAZE COMPONENT α-STABLE FITTING:RESULTS OBTAINED ON THE TVSPORTS03 CLIP

Subject	Comp. i	α_i	β_i	γ_i	δ_i
CZ	i=1	2	1	4.06	7.15
	i=2	2	1	22.44	60.82
	i=3	1.9854	1	63.99	230.31
JA	i=1	2	1	4.50	9.11
	i=2	1	1	23.37	63.89
	i=3	1.57	1	30.90	220.07
JZ	i=1	1.99	0.08	4.34	9.70
	i=2	2	-1	22.97	68.28
	i=3	1.98	1	40.07	187.77
RC	i=1	2	1	4.91	8.9
	i=2	2	1	24.88	62.69
	i=3	1.59	1	53.80	249.78
VN	i=1	1.91	1	3.35	6.58
	i=2	2	1	22.25	62.43
	i=3	1.52	1	38.85	214.20
All subjects	i=1	2	1	4.42	8.11
	i=2	2	1	23.42	63.84
	i=3	1.6	1	45.61	230.41
Ecological Sampling	i=1	2	1	3.78	9.78
	i=2	2	1	21.70	62.74
	i=3	1.76	1	59.79	245.20

TABLE II
GAZE COMPONENT α-STABLE FITTING:RESULTS OBTAINED ON THE MONICA03 CLIP

Subject	Comp. i	α_i	β_i	γ_i	δ_i
CZ	i=1	2	1	4.27	7.52
	i=2	2	1	22.44	60.82
	i=3	1.98	1	63.99	230.31
JZ	i=1	2	-1	3.60	12.40
	i=2	1.99	1	20.46	64.90
	i=3	1.75	1	30.63	197.20
NM	i=1	2	1	4.76	7.81
	i=2	1.98	1	21.32	48.8
	i=3	1.23	1	32.64	292.68
RC	i=1	1.55	1	2.68	6.92
	i=2	2	1	22.47	62.57
	i=3	1.43	1	33.50	214.15
VN	i=1	2	1	4.48	7.50
	i=2	2	1	24.15	59.05
	i=3	1.78	1	29.90	197.71
All subjects	i=1	2	1	4.47	7.54
	i=2	2	1	22.87	55.6
	i=3	1.51	1	36.69	231.06
Ecological Sampling	i=1	2	1	3.80	10.57
	i=2	2	1	22.14	58.061
	i=3	1.63	1	64.18	273.86

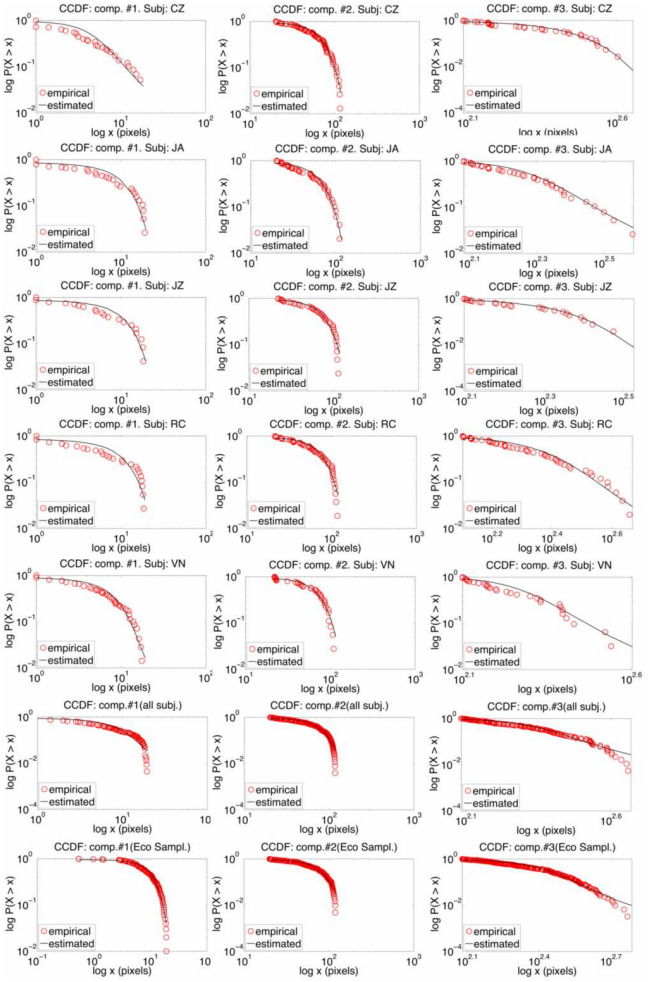

Fig. 4. Analysis of gaze shift dynamics from the tvsports03 video. From left to right, the first column shows the double log plot of the CCDF derived from the smooth-pursuit component; the center and right column, the plots related to the two saccadic components. From top to bottom, the first five rows show the CCDFs related to subjects CZ , JA, JZ, RC, VN; the sixth row presents the CCDFs obtained from the gaze magnitude distribution of all subjects. The bottom row presents the CCDF obtained from one run of the proposed algorithm.

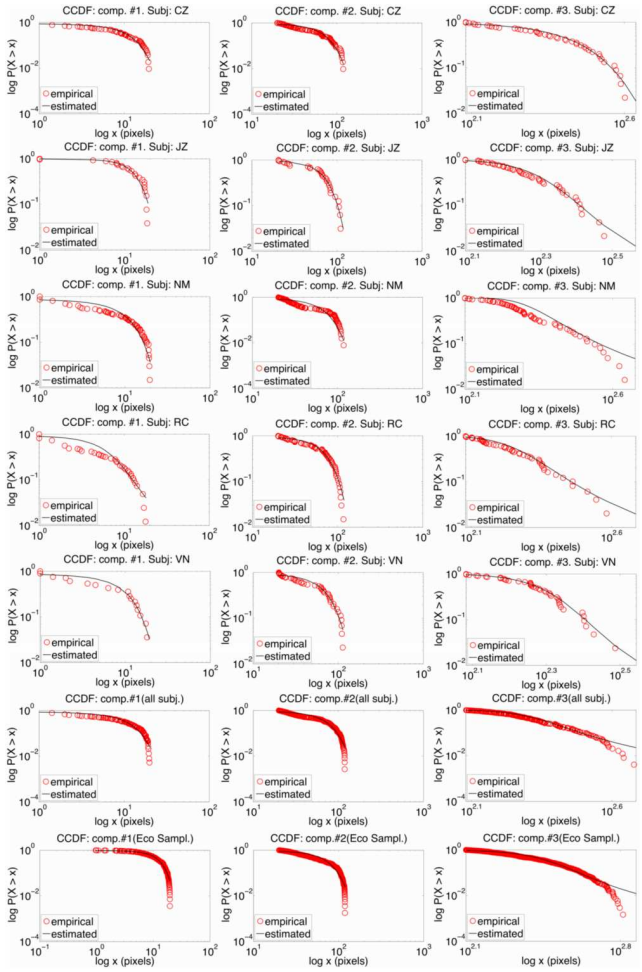

Fig. 5. Analysis of gaze shift dynamics from the monica03 video. From left to right, the first column shows the double log plot of the CCDF derived from the smooth-pursuit component; the center and right column, the plots related to the two saccadic components. From top to bottom, the first five rows show the CCDFs related to subjects CZ ,JZ, NM, RC, VN; the sixth row presents the CCDFs obtained from the gaze magnitude distribution of all subjects. The bottom row presents the CCDF obtained from one run of the proposed algorithm.

(see again Figure 3.

Having detected the spatio-temporal "gist" of the habitat, the hyperparameters of the Dirichlet distribution can be updated via (10). This is sufficient to set the bias of the "behavioral choice" (13) and the choice $z = k$ is made (14).

The actual values of the motor parameters $\eta_k = \{\alpha_k, \beta_k, \gamma_k, \delta_k\}$ corresponding to the K behaviors have been derived from the clips of the MTV-style dataset; the rationale behind this choice stems from the fact that since the latter are assembled by mixing different clips of the 'Original' dataset, parameters inferred on such clips are suitable to provide a sort of average motor behavior suitable for different types of videos.

For the examples shown here $\eta_1 = \{\alpha_1 = 2, \beta_1 = 1, \gamma_1 = 6.20, \delta_1 = 0\}$, $\eta_2 = \{\alpha_2 = 2, \beta_2 = 1, \gamma_2 = 26.10, \delta_2 = 0\}$, $\eta_3 = \{\alpha_3 = 1.72, \beta_3 = 1, \gamma_3 = 41.25, \delta_3 = 0\}$, where we have set $\delta_k = 0$, since in the sampling phase the drift is accounted for by the deterministic component of Eq. (18).

Eventually, the new FOA $\mathbf{r}_{t+1}$ is straightforwardly determined via (18). First, the drift components $-[\partial_x V, \partial_y V]^T$ are computed via (17); then, given the parameters η_k, the shift lenght components are sampled $\xi_{k,i} \sim f(\xi_{k,i}; \eta_k)$. The α-stable random vector $\boldsymbol{\xi}_k$ was sampled using the well known Chambers, Mallows, and Stuck procedure [57].

For what concerns the time sampling parameter $\tau = t_{n+1} - t_n$, $n = 0, \cdots, N$, in order to work at the frame rate of 30 fps, by assuming the time interval $T = 1$ sec and $N = 30$, the time discretization parameter is set as $\tau = T/N = 0.03$. [46]. An illustrative example, which is representative of results achieved on such data-set, is provided in Fig. 3, where the change of motor behavior regime is readily apparent as a function of the complexity of scene dynamics.

D. Computational cost

The system is currently implemented in plain MATLAB code, with no specific optimizations and running on a 2.8 GHz Intel Core 2 Duo processor, 4 GB RAM, under Mac OS X 10.5.8[1]. As regards actual performance under such setting, the average average elapsed time for the whole processing amounts to 2.175 spf (seconds per frame, frame size 640×480 pixels). More precisely, once computed the foveated frame, which takes an average elapsed time of 0.044 spf, most of the execution time is spent to compute features, 1.155 spf, and salience, 0.846 spf. The average elapsed time for obtaining patches is 0.106 spf, 0.021 spf is spent for sampling interest points, 0.001 spf is used to evaluate the complexity, and eventually 0.002 spf is used for sampling the new point of gaze. Summing up, the actual average time concerning the method proposed here, independently of feature and saliency computation (which may vary according to the technique adopted and related software and hardware optimizations), amounts to 0.130 spf. Clearly, the speed-up in this phase is due to the fact that once the set of salient interest points has been sampled, then subsequent computations only deal with N_s points in the worst case, a rather sparse representation of the original frame. For comparison purposes, the baseline algorithm [5], which is representative of the class of methods using the $\arg\max$ operation [9] for determining the gaze shift, takes an average elapsed time of 1.058 spf for the WTA computation, and 0.001 spf for the subsequent inhibition of return on the attended location. Elapsed times have been obtained using the latest version of the saliency tool box using the default parameters [9].

More generally, decision rules that boil down to the $\arg\max$ operation have $\mathrm{O}(N)$ complexity, where N is the size of the input. The original WTA procedure itself is $\mathrm{O}(N^2)$, but with specific optimization it can be reduced to $\mathrm{O}(N)$ complexity [9]. In ES the decision where to look next can be evaluated to $\mathrm{O}(N_s)$, yet $N_s \ll |L|$. Eventually, to compare with proto-object based methods that rely on the selection of the proto-object with the highest attentional weight ($\mathrm{O}(N)$, with N the number of proto-objects, e.g., [10]), the step specified by the shift equation (18) should be considered, which is $\mathrm{O}(N_V)$, $N_V \leq N_p$.

[1]In the spirit of reproducible research, the MATLAB implementation code of the ES model will be made available at http://boccignone.di.unimi.it/Ecological_Sampling.html

E. Validation

In order to verify whether the proposed model can generate statistics compared to those observed in eye-tracked subjects, we run the procedure as described above on different videos of the CRCNS 'Original' dataset[2].

The recorded FOA coordinates have been used to compute the gaze magnitude distributions. Differently from the parameter estimation stage, here we assume unlabelled distributions both for the ones obtained from ecological sampling and those composing the data-set.

Then, for each video we cluster (label) each distribution in three gaze components (smooth-pursuit and fixation + 2 saccade components) by means of VBMTS. Eventually the two samples Kolmogorov-Smirnov test is computed between each corresponding component obtained from algorithm generated and eye-tracked scanpaths considering both individual observers and the ensemble of all observers. An example of results obtained on the "tvsports03" clip, which are representative of the overall results obtained on the CNRS datase is shown in Fig. 4. It can be seen that ES generated scanpaths show strikingly similar gaze magnitude statistics described in terms of the complementary CDFs plotted on double log-scale. Table I shows the fitted α-stable component parameters for each subject participating to the experiment, the ensemble of subjects, and a scanpath generated by the ES procedure. On this clip the KS test confronting the algorithm generated and eye-tracked scanpaths fails for component 1 of subject RC (KS Statistics= 0.25836; pValue=7.4646×10^{-3}) and component 3 of subject VN (KS Statistics= 0.25032; pValue=4.8712×10^{-2}). Actually, such results are recovered when gaze shift samples from all the scanpaths, regardless of the observers, are aggregated together and used in the same distribution (row 6).

A second example is provided in Fig. 5 showing results obtained on the complex monica03 video. Table II reports the fitted α-stable parameters. In this second example the Kolmogorov-Smirnov test is not satisfied in some individual cases when the gaze component CDFs of the simulated scanpath is compared to component 1 of subjects NM (KS Statistics= 0.55742; pValue=3.3615×10^{-19}), RC (KS Statistics= 0.49375; pValue=2.8111×10^{-14}) and component 2 of subject VN (KS Statistics= 0.36991; pValue=1.2179×10^{-4}). However this is more likely to happen due to the sparsity of samples in such cases. Again, results are recovered by considering the gaze shift distribution of the observer ensemble.

It is worth noting the general trend of a nearly Gaussian behavior ($\alpha \approx 2$) of smooth pursuit / fixation(with a clear exception of subject VN) and of the first saccadic components, whilst the third component reveals a superdiffusive behavior ($\alpha < 2$). In the latter case the CCDF deviation between the empirical data and the estimated distribution that can be observed in the tail of the plot can be associated to the fact that empirical data are actually truncated (with respect to the image/field of view).

Finally, we compare the overall distributions of gaze shift amplitudes from humans, the ES model and the baseline $\arg\max$ operation [9] (Fig. 6).

To this aim we extend to videos the procedure proposed by Tatler *et al.* [2]. Note that in [2] human saccadic behavior on static images was compared against the WTA method, whereas here human amplitude distributions are derived from eye-tracking data of all subjects viewing each video. Separate simulations are run for the corresponding number of virtual observers viewing the same videos. The same time-varying saliency map is used for both ES and $\arg\max$ methods. The empirical probability densities $P(l)$ shown in Fig. 6 have been calculated from the normalized histograms of actual and simulated data. It can be seen that ES generated distributions are close to the ones exhibited by humans, whilst the distributions from the $\arg\max$ simulations fail to capture the overall heavy-tailed shapes of actual data. For the tvsports03 video (top plots) the mean, median and mode values for human and simulated data are: $mean_{Hum} = 79.73, med_{Hum} = 53.15, mode_{Hum} = 2.23$, $mean_{ES} = 65.01, med_{ES} = 47.79, mode_{ES} = 2.1$; $mean_{MAX} = 32.36, med_{MAX} = 13.89, mode_{MAX} = 2$. For the monica03 video (bottom plots) we obtained: $mean_{Hum} = 97.28, med_{Hum} = 66.94, mode_{Hum} = 1.41$; $mean_{ES} = 107.14, med_{ES} = 87.36, mode_{ES} = 1.06$; $mean_{MAX} = 36.4, med_{MAX} = 19.02, mode_{MAX} = 15$.

In particular, it can be noticed in both examples that, apart from the shorter tails, major deviations of $\arg\max$ with respect to humans (and ES) occur within the mid-range of amplitudes, which is related to complex behavior. Clearly, the slightly different trends between all distributions observed in tvsports03 and those derived from monica03 are due to the different video content.

Actually, an even more striking difference was reported in [2] between human data and the WTA simulated data. However, we must keep in mind that in [2] only static images and amplitude distributions of saccades were considered. Indeed, pictures, as opposed to natural videos, lack spatio-temporal information and thus fall short of ecological plausibility [2]. Dynamic information mitigates the limitations of using low-level saliency as the input representation since, so far, local motion features and objects/actions are often correlated [3]. This consequence is captured in Fig. 6 for small amplitude shifts, where the $\arg\max$ model exhibits a trend that is near to that of humans and ES.

V. Discussion and conclusion

In this work we have modeled a gaze shift model that allows to mimic the variability of scanpaths exhibited by human observers. The simulated behaviors are characterized by statistical properties that are close to those of subjects eye-tracked while watching complex videos. To the best of our knowledge, the ES model is novel in addressing the intrinsic stochasticity of gaze shifts and meanwhile it generalizes previous approaches proposed in the literature, [22], [33], [35], [58]–[60].

[2]This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. This includes two videos showing the foveation sequences obtained on the clips monica03 and tvsports03 from of the CRCNS 'Original' dataset and readme file. This material is 2.24 MB in size.

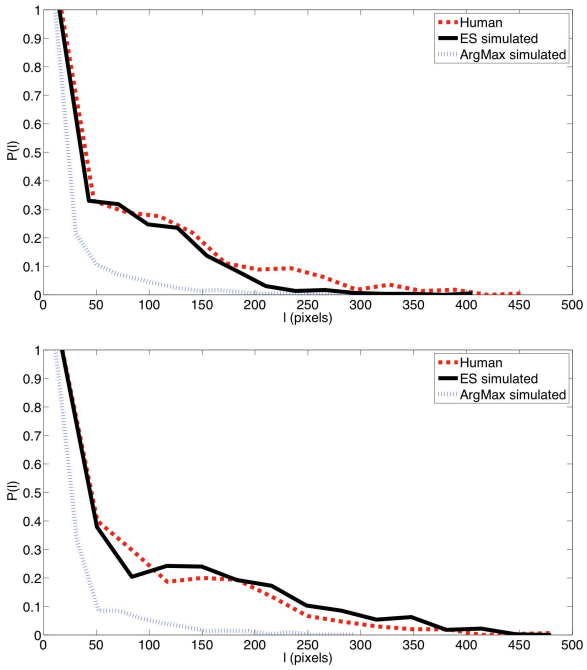

Fig. 6. Overall distributions of gaze shift amplitudes l from humans, the ES model, and the $\arg\max$ method. Top: tvsports03. Bottom:monica03 .

The core of such strategy relies upon using a mixture of α-stable motions modulated by the complexity of the scene. The strategy exploits long-tailed distributions of gaze shift lengths for the analysis of dynamic scenes, which have been usually considered limiting to static images.

The composition of random walks in terms of a mixture of α-stable components allows to treat different types of eyes movement (smooth pursuit, 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 [23], [24]. For instance, when Eq. (18) is exploited for within-patch exploration, it generates a first-order Markov process, which is compatible with most recent findings [25].

Further, this approach may be developed for a principled modeling of individual differences and departure from optimality [13] since providing cues for defining the informal notion of scanpath idiosyncrasy in terms of individual gaze shift distribution parameters. The latter represents a crucial issue both for theory [3], [19], [23] and applications [30]. Meanwhile, it stresses the importance of the role of the motor component, which is often neglected in the literature [3], [18].

One issue is how the approach presented here relates to other works in the literature. As pointed out from the beginning, scanpath variability has been abundantly overlooked in the current literature (cfr., [4]). But there are few notable exceptions. In [61] simple eye-movements patterns, in the vein of [19], are straightforwardly incorporated as a prior of a dynamic Bayesian network to guide the sequence of eye focusing positions on videos. The model presented in [62] embeds at least one parameter suitable to be tuned to obtain different saccade length distributions on static images, although statistics obtained by varying such parameter are still far from those of human data. Closer to our study is the model by Keech and Resca [63] that mimics phenomenologically the observed eye movement trajectories and where randomness is captured through a Monte Carlo selection of a particular eye movement based on its probability; probabilistic modeling of eye movement data has been also discussed in [64]. However, both models address the specific task of conjunctive visual search and are limited to static scenes. Other exceptions are given, but in the very peculiar field of eye-movements in reading [47].

The majority of models in computational vision basically resort to deterministic mechanisms to realize gaze shifts, and this has been the main route to model saccades the most random type of gaze shift [2]. Hence, if the same saliency map is provided as input, they will basically generate the same scanpath; 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 IV examples showing that the overall distributions of human and ES generated shifts on the same video are close in their statistics, see Fig. 6.

When an $\arg\max$ operation (e.g., the WTA scheme or the MAP decision rule in a probabilistic setting), the statistics of model generated scanpaths do not match those of the eye-tracked subjects and the characteristic heavy-tailed distribution of amplitudes are not recovered. This result is in agreement and extends that reported in [2].

On the other hand, models proposed in the literature that mainly focus on representational issues can be complementary to the one proposed here. Nothing prevents from using the ES gaze shift mechanism in the framework of a general top-down, object-based attention system by adopting a computation of saliency shaped in the vein of [6]. Indeed, the integration of eye guidance by interlocking ES and a full Bayesian representation of objects [6] and context [7] is the matter of ongoing research. It may be also worth noting that here eye guidance interacts with patches rather than the whole saliency map (differently from [22]). Thus, the ES model is to be naturally exploited for object-based attention schemes, relying on the notion that proto-objects drive the initial sampling of the visual scene [10], [40]. In our model, at any time t, the dynamic proto-object map is formed by the foraging eye, by considering both local and global information within the frame of the current oculomotor action. This is a possible way to account for the very notion of proto-objects as that of a ”constantly regenerating flux” advocated by Rensink [40], which makes proto-objects the bulk of interaction between perceptual and motor processes in computational models of visual attention [10].

Finally, beside theoretical relevance for modelling human behavior, the randomness of the process can be an advantage in computer vision and learning tasks. For instance, in [58] it has been reported that a stochastic attention selection mechanism (a refinement of the algorithm proposed in [33]) enables the i-Cub robot to explore its environment up to three times faster compared to the standard WTA mechanism [5]. Indeed,

stochasticity makes the robot sensitive to new signals and flexibly change its attention, which in turn enables efficient exploration of the environment as a basis for action learning [59], [60].

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. Partial support has been provided by the PASCAL2 Network of Excellence under EC grant no. 216886. This publication only reflects the authors' views.

REFERENCES

[1] M. Dorr, T. Martinetz, K. Gegenfurtner, and E. Barth, "Variability of eye movements when viewing dynamic natural scenes," *Journal of Vision*, vol. 10, no. 10, 2010.

[2] B. Tatler, M. Hayhoe, M. Land, and D. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *Journal of Vision*, vol. 11, no. 5, 2011.

[3] A. Schütz, D. Braun, and K. Gegenfurtner, "Eye movements and perception: A selective review," *Journal of Vision*, vol. 11, no. 5, 2011.

[4] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 135–207, 2013.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 1254–1259, 1998.

[6] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A Bayesian inference theory of attention," *Vision research*, vol. 50, no. 22, pp. 2233–2247, 2010.

[7] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search." *Psychological review*, vol. 113, no. 4, p. 766, 2006.

[8] M. Begum and F. Karray, "Visual attention for robotic cognition: A survey," *IEEE Trans. Autonom. Mental Dev.*, vol. 3, no. 1, pp. 92–105, 2011.

[9] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[10] M. Wischnewski, A. Belardinelli, W. Schneider, and J. Steil, "Where to Look Next? Combining Static and Dynamic Proto-objects in a TVA-based Model of Visual Attention," *Cognitive Computation*, vol. 2, no. 4, pp. 326–343, 2010.

[11] L. Elazary and L. Itti, "A bayesian model for efficient visual search and recognition," *Vision research*, vol. 50, no. 14, pp. 1338–1352, 2010.

[12] G. Boccignone, A. Marcelli, P. Napoletano, G. Di Fiore, G. Iacovoni, and S. Morsa, "Bayesian integration of face and low-level cues for foveated video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 12, pp. 1727–1740, 2008.

[13] J. Najemnik and W. Geisler, "Optimal eye movement strategies in visual search," *Nature*, vol. 434, no. 7031, pp. 387–391, 2005.

[14] A. Salah, E. Alpaydin, and L. Akarun, "A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 420–425, 2002.

[15] D. A. Chernyak and L. W. Stark, "Top–down guided eye movements," *IEEE Trans. Syst., Man, Cybern. B*, vol. 31, pp. 514–522, 2001.

[16] M. Begum, F. Karray, G. Mann, and R. Gosine, "A probabilistic model of overt visual attention for cognitive robots," *IEEE Trans. Syst., Man, Cybern. B*, vol. 40, no. 5, pp. 1305–1318, 2010.

[17] R. Canosa, "Real-world vision: Selective perception and task," *ACM Transactions on Applied Perception*, vol. 6, no. 2, p. 11, 2009.

[18] R. van Beers, "The sources of variability in saccadic eye movements," *The Journal of Neuroscience*, vol. 27, no. 33, pp. 8757–8770, 2007.

[19] B. Tatler and B. Vincent, "The prominence of behavioural biases in eye guidance," *Visual Cognition*, vol. 17, no. 6-7, pp. 1029–1054, 2009.

[20] G. Viswanathan, E. Raposo, and M. da Luz, "Lévy flights and superdiffusion in the context of biological encounters and random searches," *Physics of Life Rev.*, vol. 5, no. 3, pp. 133–150, 2008.

[21] M. Plank and A. James, "Optimal foraging: Lévy pattern or process?" *Journal of The Royal Society Interface*, vol. 5, no. 26, p. 1077, 2008.

[22] G. Boccignone and M. Ferraro, "The active sampling of gaze-shifts," in *Image Analysis and Processing ICIAP 2011*, ser. Lecture Notes in Computer Science, G. Maino and G. Foresti, Eds. Springer Berlin / Heidelberg, 2011, vol. 6978, pp. 187–196.

[23] E. Kowler, "Eye movements: The past 25 years," *Vision Research*, vol. 51, no. 13, pp. 1457–1483, 2011, 50th Anniversary Special Issue of Vision Research - Volume 2.

[24] J. Otero-Millan, X. Troncoso, S. Macknik, I. Serrano-Pedraza, and S. Martinez-Conde, "Saccades and microsaccades during visual fixation, exploration, and search: foundations for a common saccadic generator," *Journal of Vision*, vol. 8, no. 14, 2008.

[25] M. Bettenbuhl, M. Rusconi, R. Engbert, and M. Holschneider, "Bayesian selection of markov models for symbol sequences: Application to microsaccadic eye movements," *PLoS ONE*, vol. 7, no. 9, p. e43388, 2012.

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

[27] S. Ellis and L. Stark, "Statistical dependency in visual scanning," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 28, no. 4, pp. 421–438, 1986.

[28] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, September 2000.

[29] C. Harris, "On the optimal control of behaviour: a stochastic perspective," *Journal of neuroscience methods*, vol. 83, no. 1, pp. 73–88, 1998.

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

[31] E. Over, I. Hooge, B. Vlaskamp, and C. Erkelens, "Coarse-to-fine eye movement strategy in visual search," *Vision Research*, vol. 47, pp. 2272–2280, 2007.

[32] S. Siegert and R. Friedrich, "Modeling of nonlinear Lévy processes by data analysis," *Physical Review E*, vol. 64, no. 4, p. 041107, 2001.

[33] G. Boccignone and M. Ferraro, "Modelling gaze shift as a constrained random walk," *Physica A: Statistical Mechanics and its Applications*, vol. 331, no. 1-2, pp. 207–218, 2004.

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

[35] D. Brockmann and T. Geisel, "The ecology of gaze shifts," *Neurocomputing*, vol. 32, no. 1, pp. 643–650, 2000.

[36] D. Stephen, D. Mirman, J. Magnuson, and J. Dixon, "Lévy-like diffusion in eye movements during spoken-language comprehension," *Physical Review E*, vol. 79, no. 5, p. 056114, 2009.

[37] A. Reynolds, "How many animals really do the Lévy walk? Comment," *Ecology*, vol. 89, no. 8, pp. 2347–2351, 2008.

[38] H. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, pp. 1–27, 2009.

[39] D. Boyer, G. Ramos-Fernández, O. Miramontes, J. Mateos, G. Cocho, H. Larralde, H. Ramos, and F. Rojas, "Scale-free foraging by primates emerges from their interaction with a complex environment," *Proc. of the Royal Society B: Biological Sciences*, vol. 273, no. 1595, pp. 1743–1750, 2006.

[40] R. Rensink, "The dynamic representation of scenes," *Visual Cognition*, *7*, vol. 1, no. 3, pp. 17–42, 2000.

[41] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. CVPR '07*, vol. 1, 2007, pp. 1–8.

[42] T. Horowitz and J. Wolfe, "Visual search has no memory," *Nature*, vol. 394, no. 6693, pp. 575–577, 1998.

[43] M. Turner, "Landscape ecology: the effect of pattern on process," *Annual review of ecology and systematics*, vol. 20, pp. 171–197, 1989.

[44] J. Shiner, M. Davison, and P. Landsberg, "Simple measure for complexity," *Physical review E*, vol. 59, no. 2, pp. 1459–1464, 1999.

[45] M. Creutz, "Global monte carlo algorithms for many-fermion systems," *Physical Review D*, vol. 38, no. 4, p. 1228, 1988.

[46] D. Higham, "An algorithmic introduction to numerical simulation of stochastic differential equations," *SIAM review*, pp. 525–546, 2001.

[47] G. Feng, "Eye movements as time-series random variables: A stochastic model of eye movement control in reading," *Cognitive Systems Research*, vol. 7, no. 1, pp. 70–95, 2006.

[48] I. Rhee, M. Shin, S. Hong, K. Lee, S. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM Transactions on Networking,*, vol. 19, no. 3, pp. 630–643, 2011.

[49] R. Metzler and J. Klafter, "The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics," *Journal of Physics A: Mathematical and General*, vol. 37, p. R161, 2004.

[50] P. Baldi and L. Itti, "Of bits and wows: A bayesian theory of surprise with applications to attention," *Neural Networks*, vol. 23, no. 5, pp. 649–666, 2010.
[51] D. Salas-Gonzalez, E. Kuruoglu, and D. Ruiz, "Modelling with mixture of symmetric stable distributions using Gibbs sampling," *Signal Processing*, vol. 90, no. 3, pp. 774–783, 2010.
[52] C. Archambeau and M. Verleysen, "Robust bayesian clustering," *Neural Networks*, vol. 20, no. 1, pp. 129–138, 2007.
[53] I. Koutrouvelis, "Regression-type estimation of the parameters of stable laws," *Journal of the American Statistical Association*, pp. 918–928, 1980.
[54] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision research*, vol. 49, no. 10, pp. 1295–1306, 2009.
[55] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, vol. 19. Cambridge, MA: MIT Press, 2007, pp. 545–552.
[56] R. Halır and J. Flusser, "Numerically stable direct least squares fitting of ellipses," in *Proc. Int. Conf. in Central Europe on Computer Graphics, Visualization and Interactive Digital Media*, vol. 1, 1998, pp. 125–132.
[57] J. Chambers, C. Mallows, and B. Stuck, "A method for simulating stable random variables," *J. Am. Stat. Ass.*, vol. 71, no. 354, pp. 340–344, 1976.
[58] H. Martinez, M. Lungarella, and R. Pfeifer, "Stochastic Extension to the Attention-Selection System for the iCub," *University of Zurich, Tech. Rep*, 2008.
[59] Y. Nagai, "Stability and sensitivity of bottom-up visual attention for dynamic scene analysis," in *Proc. of the 2009 IEEE/RSJ Int. Conf. on Intell. Robots and Systems*. IEEE Press, 2009, pp. 5198–5203.
[60] ——, "From bottom-up visual attention to robot action learning," in *Proc. 8th Int. Conf. on Development and Learning*. IEEE Press, 2009, pp. 1–6.
[61] A. Kimura, D. Pang, T. Takeuchi, J. Yamato, and K. Kashino, "Dynamic markov random fields for stochastic modeling of visual attention," in *Proc. ICPR '08*. IEEE, 2008, pp. 1–5.
[62] T. Ho Phuoc, A. Guérin-Dugué, and N. Guyader, "A computational saliency model integrating saccade programming," in *Proc. Int. Conf. on Bio-inspired Systems and Signal Processing*, Porto, Portugal, 2009, pp. 57–64.
[63] T. Keech and L. Resca, "Eye movements in active visual search: A computable phenomenological model," *Attention, Perception, & Psychophysics*, vol. 72, no. 2, pp. 285–307, 2010.
[64] U. Rutishauser and C. Koch, "Probabilistic modeling of eye movement data during conjunction search via feature-based attention," *Journal of Vision*, vol. 7, no. 6, 2007.

Giuseppe Boccignone received the Laurea degree in theoretical physics from the University of Turin (Italy) in 1985. In 1986, he joined Olivetti Corporate Research, Ivrea, Italy. From 1990 to 1992, he served as a Chief Researcher of the Computer Vision Lab at CRIAI, Naples, Italy. From 1992 to 1994, he held a Research Consultant position at Research Labs of Bull HN, Milan, Italy, leading projects on biomedical imaging. In 1994, he joined as an Assistant Professor the Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica, University of Salerno, Italy. In 2008 he joined the Dipartimento di Informatica, University of Milan, Italy, where he currently is an Associate Professor of Perception Models, Man-Machine Interaction, Affective an Behavioral Computing. His research interests include active vision, affective computing, Bayesian models and stochastic processes for vision and cognitive science.

Mario Ferraro received the Laurea degree in theoretical physics from the University of Turin (Italy) in 1973. He has worked in Universities in England, Canada, Germany and United States, carrying on research on fuzzy sets theory, human vision, invariant pattern recognition and computational vision. Presently he is an Associate Professor of Physics at the University of Turin. His research interests include image and shape analysis, cellular biophysics and the theory of self-organising systems.