Feed and fly control of visual scanpath for foveation image processing

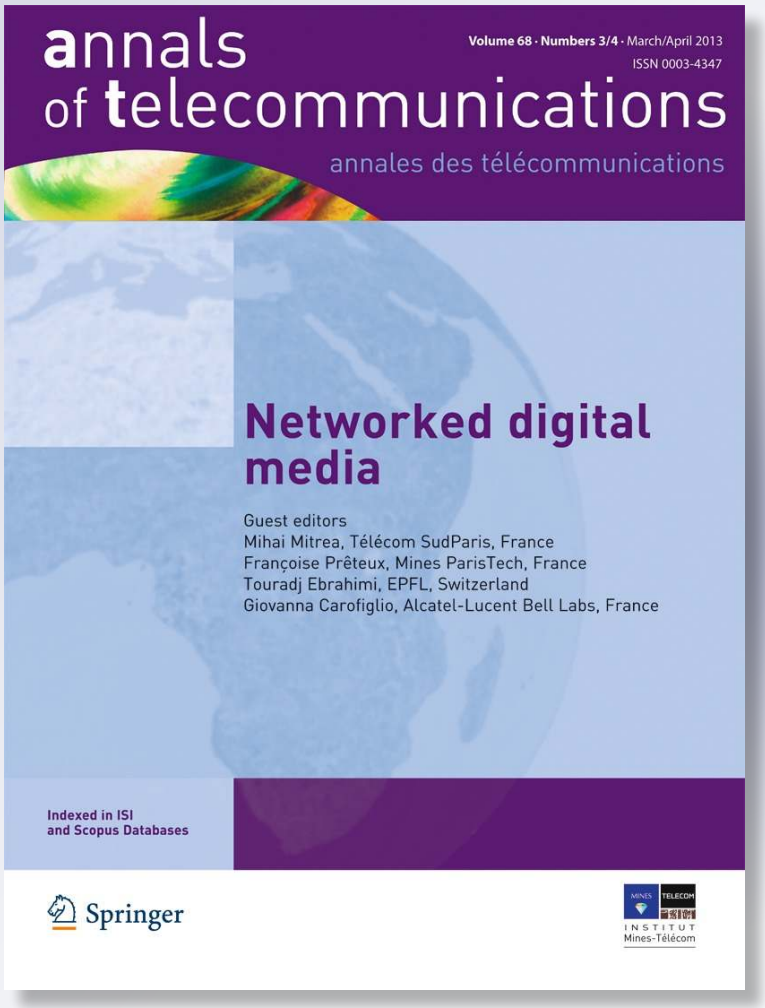
Feed and fly control of visual scanpaths for foveation image processing

Giuseppe Boccignone & Mario Ferraro

annals of telecommunications - annales des télécommunications

ISSN 0003-4347
Volume 68
Combined 3-4

Ann. Telecommun. (2013) 68:201-217
DOI 10.1007/s12243-012-0316-9



Springer

Ann. Telecommun. (2013) 68:201–217
DOI 10.1007/s12243-012-0316-9

Feed and fly control of visual scanpaths for foveation image processing

Giuseppe Boccignone · Mario Ferraro

Received: 14 January 2012 / Accepted: 4 July 2012 / Published online: 18 July 2012

Abstract Foveation-based processing and communication systems can exploit a more efficient representation of images and videos by removing or reducing visual information redundancy, provided that the sequence of foveation points, the visual scanpath, can be determined. However, one point that is neglected by the great majority of foveation models is the "noisy" variation of the random visual exploration exhibited by different observers when viewing the same scene, or even by the same subject along different trials. Here, a model for the generation and control of scanpaths that accounts for such issue is presented. In the model, the sequence of fixations and gaze shifts is controlled by a saliency-based, information foraging mechanism implemented through a dynamical system switching between two states, "feed" and "fly." Results of the simulations are compared with experimental data derived from publicly available datasets.

Keywords Eye movements · Random walk · Visual attention · Image encoding

G. Boccignone (✉)
Dipartimento di Scienze dell'Informazione, Universitá di Milano, Via Comelico 39/41, Milano, 20135 Italy
e-mail: Giuseppe.Boccignone@unimi.it

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

1 Introduction

Visual systems have a limited informational capacity [13], in the sense that only a small part of information present is registered, at any given time, and reaches levels of processing that directly influence behavior.

The human retina possesses a nonuniform spatial distribution (resolution) of photoreceptors, with highest density on that small part of the retina (about 2°–5° of visual angle) aligned with the visual axis, the fovea. The photoreceptor density rapidly decreases with distance away from the fovea; hence, the local visual frequency bandwidth also falls away. As a result, when a human observer gazes at a point in a real-world image, a variable resolution image is transmitted through the front visual channel into the high level processing units in the human brain. By contrast, traditional digital computer vision systems represent images on rectangular uniformly sampled lattices.

The motivation behind foveation image and video processing is that there exists considerable high-frequency information redundancy in the peripheral regions; thus, a much more efficient representation of images can be obtained by removing or reducing such information redundancy, bottom-up provided that foveation points (fixations) can be discovered [46].

In this perspective, visual attention plays a central role in that it controls and ensures that selected information is relevant to behavioral priorities and objectives. Kustov and Robinson have suggested that the attentional process evolved as part of the motor system [31] and eye movements are directly related to the capability of the observer for exploring the environment. In particular, the human visual system exploits saccades to actively reposition fixations on regions of

interest (the so-called focus Of attention, FOA) so as to extract detailed information from the visual environment. The succession of saccades and fixations is referred to as a *scanpath*. A scanpath of a subject scanning a natural scene is shown in Fig. 1: circular spots and lines joining spots, graphically represent, respectively, fixations and gaze shifts between subsequent fixations.

The selection of a fixation point, which allows to set the observer's FOA on the foveated region, appears to be driven by two different mechanisms: a "bottom-up" process which produces rapid scans in a saliency-driven, task-independent manner and a slower "top-down" process which is object based, task dependent, and volition controlled [39]. The degree to which these two mechanisms play a role in determining attentional selection under natural viewing conditions has been for a long time under debate [39, 41]. Certainly, top-down semantic influences do affect attentional guidance: faces and text are very attractive and are difficult to ignore, even if there is a real cost associated with looking at them [10, 11].

Thus, the possibility of realizing foveation image and video processing systems is strictly related to the capability of coping with visual attention mechanisms. The latter has gained currency in computer vision and robotics systems (see [2, 8, 18] for in-depth surveys); more recently, the efficient coding principle underlying visual attention has been exploited for image/video coding [7, 24, 32, 46] and image/video retrieval domains [4, 15]. Also, work has been done on integrating the human attention analysis into video quality assessment (see [47] for a broad survey). The rationale behind foveation coding and quality assessment is that it may not be necessary or useful to encode each image or video frame with uniform quality, since human observers will crisply perceive only a very small fraction of each frame, dependent upon their current point of fixation.

Fig. 1 Different scanpaths on a pair of images eye tracked from different human observers. *Left*, free viewing of a natural scene; *right*, natural scene embedding a face. The area of *yellow disks* marking fixations between saccades is proportional to fixation time (images from the Fixations in FAces dataset)

Despite of this flourishing interest in attention-based image and video coding systems, one important point that is neglected by the great majority of computational models (cfr. the recent review by Borji and Itti [8]), is the "noisy" variation of the exploration exhibited by different observers when viewing the same scene. Indeed, though some particular locations in the image attract the gaze of different observers (and might be predicted by bottom-up or top-down visual attention models), the moment-to-moment relocation of gaze is different among observers or even by the same subject along different trials [30, 39]. This peculiar characteristic can be appreciated by considering Fig. 1. Such random variation in individual scanpaths (with regard to chosen fixations, spatial scanning order, and fixation duration) still holds even when the image contains semantically rich "objects" (cfr. Fig. 1, images on the right).

The variability of saccades is interesting because a number of prior studies have shown that it occurs ubiquitously, and it may mediate a variety of saccadic and perceptual phenomena. 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 sensing, information processing, movement planning, and executing [1]. At this level, the issue of stochasticity in scanpaths, debated in early studies [17, 37], 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" [21].

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. Systematic tendencies in oculomotor behavior can be thought of as regularities that are common across all instances of and manipulations to the behavior. Such tendencies can be seen, for instance, in saccade amplitudes, which show a positively skewed, long-tailed distribution in most experimental settings in which complex scenes are viewed [43]. Under certain conditions, these can provide a signature of

the oculomotor behavior peculiar to an individual (the idiosyncrasy of scanpaths [35]).

In a different perspective, by analyzing the spatial pattern of gaze shifts—local exploration followed by long shifts, see Fig. 2—Brockmann and Geisel [9] have shown that a visual system producing Lévy flights implements a more efficient strategy of shifting gaze in a random visual environment than any strategy employing a typical scale in gaze shift magnitudes; evidence of Lévy diffusive behavior of scanpath has been presented in [42]. Indeed, such behavior gives rise to saccade amplitude distributions of the kind discussed by Tatler and Vincent [43].

Building upon [9], in [5], a gaze-shift model (the Constrained Lévy Exploration, CLE) has been proposed. Such model is somehow akin to models of simple animal foraging, where the visual system hunts for areas that are rich in visual saliency, under the assumption that eye movements and animal foraging address in some way the problem of searching randomly distributed sites whose exact locations are not known a priori. Under the foraging metaphor, the eye (and the brain modules controlling the eye behavior) is the forager, the visual saliency surface is the foraging landscape, points of fixations are foraging sites, and saccades are flights from one site to another.

In [5], eye gaze shifts are modeled by Lévy flights, constrained by a potential which is a function of the saliency (landscape). Lévy flights, as opposed, for instance, to usual random walk, may be essential for optimal search in foraging, where optimality is related to efficiency; that is, the ratio of the number of sites visited to the total distance traversed by the forager [45]. The model, while accounting for scanpath randomness, roughly mimicked a straight reactive behavior of the observer/forager with respect to the potential designed on the basis of landscape saliency. In other terms, it represented a low-level layer of a complex sensorimotor control module.

However, one could argue, from an evolutionary standpoint, that specific search mechanisms could have been subsequently learned and "wired" in order to improve the exploration reliability and efficiency. For example, it has been suggested [16] that to optimize the search of the target sites, locomotion rules need to be embedded within the search mechanism.

Thus, in [6], a model has been presented where the process of random search can, under certain conditions on the saliency of the image, be overruled by a simple local deterministic rule, resulting in a hybrid dynamical system (hybrid constrained search, HCS). Such process can be seen as the result of the action of a higher-level control system superimposed to the lower stochastic one. This idea is consistent with view, dating back to Jackson's work [27], that the evolution of the nervous system can be conceived as an incremental process in which higher level control systems overrule lower levels.

The results presented in [6] make clear that the addition of deterministic rules results in more efficient and robust processes of visual exploration. In this sense, the layered organization of the HCS model provides a better model of human gaze-shift behavior than CLE, in that humans appear able to perform an efficient scanpath under different environmental conditions.

Here, we extend the basic insight of HCS by taking into account some issues which are critical in the characterization and control of scanpath generation such as the degree of information about the scanned scene available to the observer, fixation duration, inhibition of return (IOR) to the same fixation point. In the following, it will be shown how the HCS model can be extended so to embed such issues yielding to the informed HCS model (IHCS).

2 Background

Consider a random walker moving under the influence of an external force; at time t, the transition from the current position $\mathbf{r}(t)$ to a new position $\mathbf{r}_{\text{new}}(t)$, $\mathbf{r}(t) \to \mathbf{r}_{\text{new}}(t)$, is given by

$$\mathbf{r}_{\text{new}}(t) = \mathbf{r}(t) + \mathbf{g}(\mathbf{r}(t)) + \boldsymbol{\eta}. \tag{1}$$

The trajectory of the variable $\mathbf{r}$ is determined by a deterministic part $\mathbf{g}$, the drift, and a stochastic part $\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a random vector with components

$$\eta_x = l\cos(\theta), \qquad \eta_y = l\sin(\theta), \tag{2}$$

where the angle θ represents the flight direction and $l = |\boldsymbol{\eta}|$ is the jump length.

If a uniform distribution of directions is assumed, then, the walker's motion is determined by the probability density function f from which amplitude l is sampled, $l \sim f$. For instance, if f is a Gaussian distribution, the usual Brownian motion occurs.

However, Brownian motion is nothing but a special case within the family of stochastic processes qualifying as natural models for random noise sources. Other types of motion can be generated by resorting to the class of the so-called α-stable distributions [19]. These form a four-parameter family of continuous probability densities, say $f(x;\alpha,\beta,\gamma,\delta)$, parametrized by the skewness β (measure of asymmetry), scale γ (width of the distribution) and location parameters δ and, most

important, the *characteristic exponent* α or index of the distribution that specifies the asymptotic behavior of the distribution.

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

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

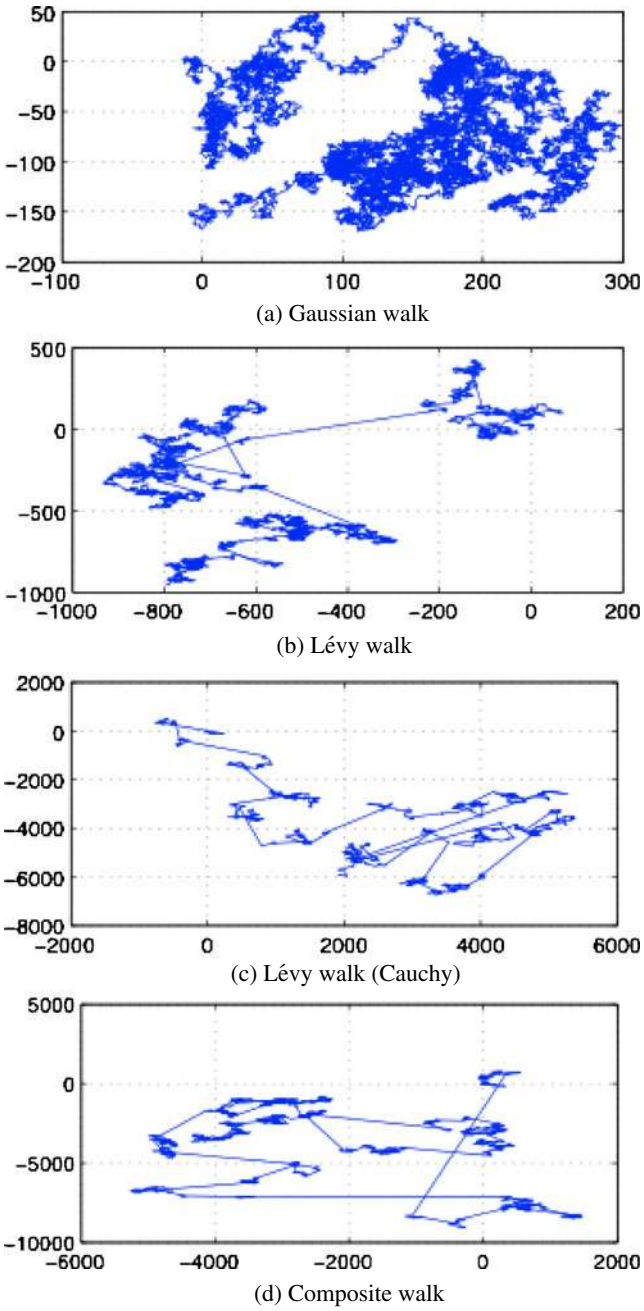
The first expression holding if $\alpha \neq 1$ and the second if $\alpha = 1$.

Special cases of stable distributions whose pdf can be written analytically are given for $\alpha = 2$, the normal distribution $f(x;2,0,\gamma,\delta)$, for $\alpha = 1$, the Cauchy distribution $f(x;1,0,\gamma,\delta)$, and for $\alpha = 0.5$, the Lévy distribution $f(x;0.5,1,\gamma,\delta)$; for all other cases, only the characteristic function is available in closed form, and numerical approximation techniques must be adopted for both sampling and parameter estimation [12, 29, 34].

When stable distributions are used to characterize the step lengths l of a random walker as given by Eq. 1, since f scales, asymptotically, as $l^{-1-\alpha}$, then relatively long jumps are more likely when α is small. In fact, by sampling $l \sim f(l;\alpha,\beta,\gamma,\delta)$, for $\alpha \geq 2$, the usual random walk (Brownian motion) occurs; if $\alpha < 2$, the distribution of jump lengths is "broad" and the so-called Levy flights take place.

Examples of Lévy flights, obtained from Eq. 1 with no drift ($\mathbf{g} = 0$), are presented in Figs. 2b, c: these typically exhibit local walk interleaved with long jumps and should be compared to Brownian motion plot in Fig. 2a. The bottom plot illustrates a random walk pattern obtained as a composite process simulated by sampling the step length from a mixture of two α-stable distributions indexed by $\alpha_1 = 2$ and $\alpha_2 = 1$, respectively, and mixture weights $w_1 = 0.4$ and $w_2 = 0.6$. It is worth noting in the latter case that the walking pattern could be identified as a Lévy pattern though, in contrast with the other cases, the pattern is composite (Brownian and Cauchy).

The general applicability of Lévy flights in ecology and biological sciences is still open to debate, as recent experimental data show that the movement patterns of various marine predators and terrestrial animal exhibit a Lévy walk pattern in areas with low abundance of preys or foods and Brownian walk pattern (a sort of food tracking) in areas with high abundance [14].



(a) Gaussian walk

(b) Lévy walk

(c) Lévy walk (Cauchy)

(d) Composite walk

Fig. 2 Different random walks obtained by sampling lengths l for different α parameters; the walks shown have been generated setting $\alpha = 2$ in plot (Fig. 2a), $\alpha = 1.6$ in plot (Fig. 2b), $\alpha = 1$ in plot (Fig. 2c); *bottom plot* (Fig. 2d) represents a composite walk sampled from a mixture of two stable distributions indexed by $\alpha = 2$ and $\alpha = 1$ parameters

Foraging patterns obtained through a composite strategy have gained currency in the literature as being the most effective from a behavioral and evolutionary standpoint. Indeed, in complex environments, optimal searches should result from a mixed/composite strategy (generating patterns similar to that of Fig. 2d), different

kinds of motion can be adopted depending on the structure of the landscape in which the organism moves [36, 38]. In order to account for a composite strategy in a simple and efficient way, the HCS model was introduced [6] whose basic features can be summarized as follows.

Assume as input a color image. Each color image is a vector, $\mathbf{I}$, that is a map from the support $\Omega \subseteq R^2$ to an m-dimensional range, $\mathbf{I} : \Omega \to C \subseteq R^m$.

A saliency field s upon the image is a landscape upon which the visual exploration is performed. Formally, s is a scalar field obtained through a transformation $\mathbf{I} \mapsto s(\mathbf{I}) \in \mathbf{R}$ (see [8] for different ways of defining the mapping).

Let $\mathbf{r}(t) \in \Omega$ be the current position of the gaze, and let ρ be an arbitrary positive number; define $\mathbf{r}^*(t)$ as

$$\mathbf{r}^*(t) = \arg\max_{\mathbf{r}'(t)} \{s(\mathbf{r}'(t))\}_{\mathbf{r}'(t) \in \mathcal{N}_{\mathbf{r}(t)}}, \tag{3}$$

where $\mathcal{N}_{\mathbf{r}(t)}$ is the circle of radius ρ centered on $\mathbf{r}(t)$ and $\mathbf{r}(t) \neq \mathbf{r}'(t)$, and let $\Delta s = s(\mathbf{r}^*(t)) - s(\mathbf{r}(t))$.

The HCS model determines the next position $\mathbf{r}_{\text{new}}(t)$ of the gaze, computed at time t, as follows.

Let $\nu > 0$ be an arbitrary threshold and $\boldsymbol{\eta}$ a stochastic vector with components $(\eta_x, \eta_y)^T$ defined as in Eq. 2. Finally, consider a potential V as a time-varying scalar function of the saliency

$$V(x, y, t) = \exp(-\tau_V s(x, y, t)) \tag{4}$$

where τ_V is a damping parameter.

Then, the next position $\mathbf{r}_{\text{new}}(t)$ is given by developing Eq. 1 in the following system of equations:

$$\mathbf{r}_{\text{new}}(t) = \xi \mathbf{r}^*(t) + (1 - \xi)\,[\mathbf{r}(t) - \nabla V + \boldsymbol{\eta}] \tag{5a}$$

$$\xi = H\,(\Delta s - \nu) \tag{5b}$$

Here ,∇V is the gradient of V and H is the Heaviside function.

If $\Delta s > \nu$, then $\xi = 1$ and the foraging eye is in the *intensive stage* : the gaze moves directly to $\mathbf{r}_{\text{new}}(t) = \mathbf{r}^*(t)$; in other words, if there exist candidate target sites within a "direct vision" distance ρ with associated an increase of saliency large enough, the visual system carries out a deterministic search selecting the one with the largest saliency.

On the other hand, if $\xi = 0$, the *extensive stage* is performed and Eq. 5a becomes

$$\mathbf{r}_{\text{new}}(t) = \mathbf{r}(t) - \nabla V + \boldsymbol{\eta}, \tag{6}$$

showing that the new gaze position is determined by: a) $-\nabla V$, the force field shaped by the saliency landscape; b) the stochastic vector $\boldsymbol{\eta}$.

Direction and length of the random vector $\boldsymbol{\eta}$ are sampled from the uniform and α-stable distribution, respectively:

$$\theta \sim Unif(0, 2\pi), \tag{7}$$

$$l \sim \varphi(s)\, f(l; \alpha, \beta, \gamma, \delta). \tag{8}$$

In [6], following [5] and [9], symmetric Cauchy flights ($\alpha = 1, \beta = 0$) have been exploited.

Along the extensive stage, θ and l summarize the internal action choice of the forager and the function $\varphi(s)$ modifies the pure Lévy flight, since the probability to move from a site to the next site depends on the "strength" of a bond

$$\varphi(s) = \frac{\exp(-\beta_P(s(\mathbf{r}(t)) - s(\mathbf{r}_{\text{new}}(t))))}{\sum_{\mathbf{r}'_{\text{new}}} \exp(-\beta_P(s(\mathbf{r}(t)) - s(\mathbf{r}'_{\text{new}}(t))))} \tag{9}$$

that exists between them, and β_P being a parameter modulating such strength.

It should be remarked that the stochastic process underlying long gaze shifts should be in principle subdivided in two steps: flight proposal and acceptance of the flight; these two steps together provide an approximation of a highly complex sensory-motor process, which is far from being fully understood [39]. In this perspective, the plausible center of a new fixation $\mathbf{r}_{\text{new}}(t)$, should be eventually accepted on the basis of some decision function $\mathscr{D}(\mathbf{r}_{\text{new}}(t), T)$, where T is a parameter or a set of parameters akin to summarize the "readiness" of the forager to engage in the flight. Clearly, this is a complex issue to take into account and encompasses subtleties that are far beyond the scope of this paper. A simplified decision rule is to evaluate the jump proposal $\mathbf{r}(t) \to \mathbf{r}(t)_{\text{new}}$ through an acceptance process, implemented by a Metropolis algorithm [6]: the the target site $\mathbf{r}_{\text{new}}(t)$ is accepted with probability

$$p(a|\mathbf{r}(t)_{\text{new}}, \mathbf{r}(t)) = \min\{1, \quad \exp(\Delta s/T)\}, \tag{10}$$

where a is a binary random variable ($a = 1$, acceptance, $a = 0$ rejection). Such probability depends on the gain of saliency and on a "temperature" T. The values of T determine the amount of randomness in scanpath generation, and the role of this parameter has been extensively discussed in [6].

Finally, if no suitable candidate FOA $\mathbf{r}(t)_{\text{new}}$ has been determined during either the intensive or extensive stage, the current fixation point $\mathbf{r}(t)$ is retained.

Although, the layered organization of the HCS system provides a better model of human gaze-shift behavior than CLE, yet some issues that are crucial for

modeling scanpaths on images, and in turn visual attention, are only implicitly considered or overlooked.

First, the switch between intensive and extensive search based on thresholding, Eq. 5b, is a rather rough solution. The decision to stay in one state or the other may depend by several factors: internal state of the foraging eye, waiting (fixation) time, and general appearance of the landscape (or related fluctuations, in case of a time-varying landscape such as that generated in videos). Indeed, the intensive stage can be interpreted in terms of visual fixation. Yet, a fixation is not simply the maintenance of the visual gaze on a single location but rather a slow oscillation of the eye [30]. They are never perfectly steady and different mechanisms can be at their origin, e.g., microsaccades. One possible function for microsaccades is to bring the line of sight to a succession of locations of interest, functioning as a search or scan pattern, analogous to the function of larger saccades. Thus, eye fixations are better defined as the amount of continuous time spent looking within a circumscribed region (e.g., minimum 50 ms within a spatially limited region, typically 0.5°–2.0° of visual angle [22]). To account for all such complex factors, a probabilistic mechanisms could be more suitable and flexible.

Second, the kind of information foraging performed by the eye, especially on static images, is a sort of foraging with depletion of visited sites (destructive foraging) [28]: speed and accuracy with which a site or an object is detected are first briefly enhanced (for perhaps 100–300 ms) after the object is attended, and then, detection speed and accuracy are impaired (for perhaps 500–3,000 ms); this is well known as the IOR mechanism, which promotes exploration of new, previously unattended loci in the scene during visual search or foraging by preventing attention from returning to already-attended sites. Clearly, the amount of depletion in a circumscribed region is related to the previously discussed fixation time issue.

Third, the total amount of information about the visual landscape available to the forager can influence the generated scanpath or foraging pattern. Total absence of information or full information gives rise to different scanpaths. In practical terms, by fixing the landscape, the amount of information may also simply depend on the viewing distance display dimension: looking at a picture or a video on a cell phone is different than looking on 40 in. TV. Thus, a scanpath generation model should provide some control on this point.

In the following section, we present how such issues can be taken into account by the extended version of the HCS model, namely, the informed HCS.

3 The IHCS model

Rewite Eq. 5a more generally as

$$\mathbf{r}_{\text{new}}(t) = \mathbf{r}(t) + \xi \mathbf{g}_1(\mathbf{r}(t)) + (1-\xi)\left[\mathbf{g}_2(\mathbf{r}(t)) + \boldsymbol{\eta}\right], \quad (11)$$

where $\mathbf{g}_2 = -\nabla V$ and the drift term $\mathbf{g}_1$ will be discussed later.

When $\xi = 1$, the forager is engaged in an intensive search, while depleting the sites, he visits; denote such state "feed." When $\xi = 0$ the forager performs extensive search, a state denoted "Fly". The dynamics of the system can be described in terms of a stochastic machine (a probabilistic finite-state machine) as represented in Fig. 3, whose behavior is detailed in the following paragraph.

Feed or fly switching Let q be the probability of remaining in the feeding state, while the transition $\xi = 1 \rightarrow \xi = 0$ occurs with probability $1-q$. Then, we can use q as the parameter of the Bernoulli distribution, $Bern(\xi; q) = q^{\xi}(1-q)^{1-\xi}$ for $\xi \in \{0, 1\}$.

This way, the choice of action, keep feeding ($\xi = 1$) or engage in a flight ($\xi = 0$), can be conceived as a decision sampled from $Bern(\xi; q)$ with probability $p(\xi = 1) = q$ or $p(\xi = 0) = 1 - q$.

On the other hand, it is clear that the bias of such "coin tossing" procedure is time and space dependent. In order to account for the dependency of the fixation time on the information (saliency) contained in the direct vision range of the current FOA, we allow q to depend on the number of visited interest points (food items) present in this FOA "patch" (space dependency). Formally, we model q with an exponential function,

$$q = \exp\left(-\frac{n_s(t)}{\mu}\right), \quad (12)$$

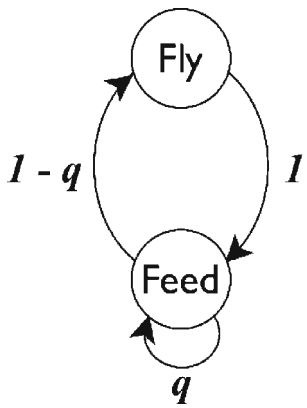


Fig. 3 The probabilistic finite-state machine representing the stochastic forager

where $n_s(t)$ is the number of saliency points locally explored at time t and μ represents the mean feeding rate of the forager.

Eventually, the transition $\xi = 0 \rightarrow \xi = 1$ occurs with probability 1; namely, after a flight, the foraging eye always is prompted to engage in the intensive stage. This in principle does not imply that local search be always actually performed: if conditions for feeding are not met and/or because of the randomness of the process, the transition $\xi = 1 \rightarrow \xi = 0$ may occur before such stage takes place, as detailed in the sequel.

Intensive search with site depletion Coming back to Eq. 11, the drift term $\mathbf{g}_1$ applied in the Feed state is modeled in order to account for the external force exerted by salient points within the FOA patch of current fixation $\mathbf{r}(t)$. Set $\mathbf{g}_1 = -\nabla U$. In a foraging framework, animals are expected to be attracted or repelled from certain food sources (interest sites); therefore, $U(\mathbf{r}, t)$ 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. More precisely, $U(\mathbf{r}, t) = \psi(|\mathbf{r}^*(t) - \mathbf{r}(t)|^2)$ for some function ψ. Define, for simplicity, ψ as the identity function $\psi(|\mathbf{r}^*(t) - \mathbf{r}(t)|^2) = |\mathbf{r}^*(t) - \mathbf{r}(t)|^2$ where $\mathbf{r}^*(t)$ belongs to a set of N_U sites selected within the FOA patch according to some rule, e.g, the top-N_U most attractive items in terms of saliency, or randomly sampled. Then, at each time step, the gradient of the potential can be obtained

$$-\nabla U(\mathbf{r}, t) = -2(\mathbf{r}(t) - \mathbf{r}^*(t)). \tag{13}$$

In ecology, this setting is adopted to model an animal attracted to the point $\mathbf{r}^*$ such as a food site

Under these assumptions, when the foraging eye is in Feed state ($\xi = 1$), Eq. 11 becomes

$$\mathbf{r}_{\text{new}}(t) = \mathbf{r}(t) - 2(\mathbf{r}(t) - \mathbf{r}^*(t)), \tag{14}$$

where points $\mathbf{r}^*(t)$ plays the role of local attractors.

Further, when any site $\mathbf{r}_{\text{new}}(t)$ is reached, destructive foraging is performed in a small neighborhood $\mathcal{N}_\varepsilon(\mathbf{r}_{\text{new}}(t))$:

$$s_{\text{new}}(\mathcal{N}_\varepsilon(\mathbf{r}_{\text{new}}(t))) = k_d s(\mathcal{N}_\varepsilon(\mathbf{r}_{\text{new}}(t))), \tag{15}$$

where k_d is a depletion constant in the range [0, 1].

Summing up, when the foraging eye is engaged in the intensive stage of local search and feeding, at each time step t, Eqs. 14 and 15 are computed, the number of visited sites is incremented, $n_s(t) \leftarrow n_s(t) + 1$, and the parameter q is computed according to Eq. 12; eventually, the choice of the next action, feed or fly, is sampled:

$$\xi \sim Bern(\xi; q). \tag{16}$$

Note that the probability $1 - q$ of leaving the feed state strictly depends on the food intake $n_s(t)$ (and on the mean feeding rate μ of the forager); however, the decision to stay or to leave is sampled from Eq. 16. Thus, it may occur earlier or later with respect to its expected time. This provides a simple way to account for statistical variability in fixation time at a given spatial location [30].

Information guided Lévy flights How does information gathered at the preattentive stage and from peripheral (extrafoveal) regions of the retina influence the generation of scanpaths? In visibility models of saccadic eye movements [30], rather than taking the line of sight to a region that already stands out from the neighboring surround, each saccade is directed to the location that would yield the highest probability of finding the target. A search strategy can be seen as one of sending the line of sight to locations that maximized search performance by considering, before each saccade, the effect of the eyes' next landing position on the visibility of all locations throughout the visual field.

Hence, under the assumption that a successful action requires the capacity of predicting the expected consequences of action, the pair (θ, l) is chosen in order to maximize the posterior distribution $p(\theta, l|s(\mathbf{r}_{\text{new}}(t)), s(\mathbf{r}(t)), \mathbf{r}_{\text{new}}(t), \mathbf{r}(t))$, namely,

$$(\theta^*, l^*) = \arg\max_{\theta, l} p(\theta, l|s(\mathbf{r}_{\text{new}}(t)), s(\mathbf{r}_{\text{new}}(t)), \mathbf{r}_{\text{new}}(t), \mathbf{r}(t)). \tag{17}$$

The selection of action parameters should be conditioned on the gain achievable by shifting to a new information state $(\mathbf{r}_{\text{new}}(t), s(\mathbf{r}_{\text{new}}(t)))$ from the current state $(\mathbf{r}(t), s(\mathbf{r}(t)))$ and can be formalized in terms of a probabilistic generative model.

Define the joint probability

$$p(\theta, l, s(\mathbf{r}_{\text{new}}(t)), s(\mathbf{r}(t)), \mathbf{r}_{\text{new}}(t), \mathbf{r}(t)).$$

The latter can be factorized as

$$\begin{aligned} &p(\theta, l, s(\mathbf{r}_{\text{new}}(t)), s(\mathbf{r}(t)), \mathbf{r}_{\text{new}}(t), \mathbf{r}(t)) \\ &\quad = p(s(\mathbf{r}_{\text{new}}(t))|s(\mathbf{r}(t)), \mathbf{r}_{\text{new}}(t), \mathbf{r}(t), \theta, l) \\ &\qquad \times p(\mathbf{r}_{\text{new}}(t)|\mathbf{r}(t), s(\mathbf{r}(t)), \theta, l) \\ &\qquad \times p(\mathbf{r}(t), s(\mathbf{r}(t))|\theta, l) \\ &\qquad \times p(\theta, l). \end{aligned} \tag{18}$$

Such factorization can be explained as follows.

The first factor in Eq. 18 provides the likelihood of jumping at any site $\mathbf{r}_{\text{new}}(t)$, starting from current position $\mathbf{r}(t)$ (the current FOA) evaluated in terms of

saliency gain and does not depend on θ, l. The second factor represents the motor action responsible for the gaze shift, that is the probability of the shift $\mathbf{r}(t) \rightarrow \mathbf{r}_{\text{new}}(t)$ given a pair (θ, l), and does not depend on current saliency $s(\mathbf{r}_(t))$. The third factor stands for the joint density of location and saliency and does not depend on (θ, l), thus $p(\mathbf{r}(t), s(\mathbf{r}(t))|\theta, l) = p(\mathbf{r}(t), s(\mathbf{r}(t)))$; further, since for the purposes of the present work we are not making prior assumptions about any location $\mathbf{r}(t)$ being more likely to be a salient point than other locations $\mathbf{r}^{'}(t)$ (as opposed, for instance, to context-based models [8]), we set $p(\mathbf{r}(t), s(\mathbf{r}(t))) = const$. The latter term is the joint prior probability on saccade amplitude and directions, which we assume as independently distributed. Hence, Eq. 18 can be approximated as

$$\begin{aligned} &p(\theta, l, s(\mathbf{r}_{\text{new}}(t)), s(\mathbf{r}_(t)), \mathbf{r}_{\text{new}}(t), \mathbf{r}(t)) \\ &\quad \approx p(s(\mathbf{r}_{\text{new}}(t))|s(\mathbf{r}(t)), \mathbf{r}_{\text{new}}(t), \mathbf{r}(t)) \\ &\quad\quad \times p(\mathbf{r}_{\text{new}}(t)|\mathbf{r}(t), \theta, l)\, p(\theta)\, p(l). \end{aligned} \tag{19}$$

In Eq. 19, the likelihood of jumping at a certain site $\mathbf{r}_{\text{new}}(t)$, starting from current FOA $\mathbf{r}(t)$, can be evaluated as

$$\begin{aligned} &p(s(\mathbf{r}_{\text{new}}(t))|s(\mathbf{r}(t)), \mathbf{r}_{\text{new}}(t), \mathbf{r}(t)) \\ &\quad = \frac{\exp(-\beta_P(s(\mathbf{r}(t)) - s(\mathbf{r}_{\text{new}}(t))))}{\sum_{\mathbf{r}^{'}_{\text{new}}} \exp(-\beta_P(s(\mathbf{r}(t)) - s(\mathbf{r}^{'}_{\text{new}}(t))))}. \end{aligned} \tag{20}$$

In other terms, the likelihood modifies the pure Lévy flight, in that the jump has a higher probability to occur if the target site is strongly connected in terms of saliency to the current one, similarly to Eq. 9

The remaining factors in Eq. 19 summarize the motor action $p(\mathbf{r}_{\text{new}}(t)|\mathbf{r}(t), \theta, l)$ and the prior probabilities on action parameters, $p(\theta)$ and $p(l)$.

The prior distribution of flight directions can be taken as the uniform distribution in the $[0, 2\pi]$ interval, $Unif(0, 2\pi)$; the prior distribution of jump lengths is taken to be an instance of the family of α-stable distributions $f(l; \alpha, \beta, \gamma, \delta)$. In such framework, the parameters of the distribution can be considered as akin to "internal" motor parameters. Note that if θ and l were straightforwardly sampled from the priors $p(\theta)$ and $p(l)$, respectively, and inserted in Eq. 2, a classic Lévy flight driven by external potential would occur.

Eventually, by using Bayes' rule and Eq. 19, the choice of action parameters (Eq. 17) can be written as

$$\begin{aligned} &\arg\max_{\theta,l} p(\theta, l|s(\mathbf{r}_{\text{new}}(t)), s(\mathbf{r}(t)), \mathbf{r}_{\text{new}}(t), \mathbf{r}(t)) \\ &\quad \approx \arg\max_{\theta,l} p(s(\mathbf{r}_{\text{new}}(t))|s(\mathbf{r}(t)), \mathbf{r}_{\text{new}}(t), \mathbf{r}(t)) \\ &\quad\quad \times p(\mathbf{r}_{\text{new}}(t)|\mathbf{r}(t), \theta, l)\, p(\theta)\, p(l). \end{aligned} \tag{21}$$

Equation 21 can be evaluated by: (1) sampling candidate gaze shifts using the prior $p(\mathbf{r}_{\text{new}}(t)|\mathbf{r}(t), \theta, l)\, p(\theta)\, p(l)$ and (2) weighting the samples through the likelihood specified in Eq. 20.

Prior sampling can be accomplished through simple ancestral sampling [3] on the probabilistic graphical model tying random variables $\theta, l, \mathbf{r}(t), \mathbf{r}_{\text{new}}$, namely, the directed graph $\{\theta, l, \mathbf{r}(t)\} \mapsto \mathbf{r}_{\text{new}}(t)$, where $\mapsto$ is the edge denoting conditional dependency between random variables (nodes); such procedure amounts to forward sampling from ancestor or parent nodes of $\mathbf{r}_{\text{new}}$, $pa\,(\mathbf{r}_{\text{new}}) = \{\theta, l, \mathbf{r}(t)\}$ to the descendant node $\mathbf{r}_{\text{new}}$:

$$\theta_k \sim Unif(0, 2\pi), k = 1, \cdots, K \tag{22}$$

$$l_k \sim f(l_k; \alpha, \beta, \gamma, \delta) \tag{23}$$

$$\mathbf{r}_{new,k}(t) \sim p(\mathbf{r}_{\text{new}}(t)|\mathbf{r}(t), \theta_k, l_k). \tag{24}$$

where K is the number of samples.

Note that together, Eqs. 22–24 provide a set of K motor actions that a priori could be undertaken by the forager. In particular, Eq. 24, since representing the shift $\mathbf{r}(t) \rightarrow \mathbf{r}_{\text{new}}(t)$, can be implemented via Eq. 6 (in the molecular dynamics literature, this approach is known as Langevin Monte Carlo [33]). In other terms, the motor step specified through Eq. 6 is used at this stage as an internal model to simulate possible candidate flights among which the most likely actual flight is eventually determined by selecting the most suitable flight parameters by Eq. 21.

The critical parameter here is the number K of samples generated, which can be directly interpreted as related to the information available to the observer: in the limit of K equal to the dimension of the image support, we are in the case of full information or full visibility.

4 Simulation

The goal of experiments described here is twofold: on the one hand, we wanted to quantitatively compare IHCS against HCS with regard to exploration performance and on the other hand, a qualitative comparison of IHCS-generated scanpaths with scanpaths eye tracked from human subjects was taken into account in order to assess the capability of IHCS to mimic human gaze behavior on images containing either low-level cues and semantically relevant objects (faces).

4.1 Datasets

Hou and Zhang CVPR07 dataset The dataset is downloadable at http://www.klab.caltech.edu/~xhou/projects/spectralResidual/spectralresidual.html. This is a collection of 62 natural scene images. Further, such images were provided to four naive subjects. Each subject was instructed to select regions where objects are presented; if each of the subject reported impossible to define an object in a certain image, that image would be rejected from the dataset. Note that the purpose of the experiment was different from segmentation; the main concern in segmentation tasks is the abrupt changes in space. In building up the dataset, hand labelers concentrated only on the edges between the foreground and the background, so to suggest a set of candidate proto-objects. For details refer to [23].

Cerf 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 colors, races, etc. Each image has a corresponding background image with no faces for comparison. The data include the fixations recorded via eye tracking of eight subjects (see [10, 11], for details). In addition to fixation data, an annotation of the entire dataset is provided, where the location and labeling of faces in images are given.

4.2 Implementation details

The exploration of the visual field performed according to the rules of selection described above can be summarized in the *informed hybrid constrained search algorithm*:

The input of the IHCS algorithm is the salience map s computed from image $\mathbf{I}$, the desired number of fixations N_{fix}, and a set of parameters. The values of the k_d, μ parameters have been derived via ROC analysis of results obtained from a preliminary trial of experiment 1 described below by using a small subset of 10 images randomly chosen from the CVPR07 dataset. The settings of the remaining parameters are discussed in detail in the sequel.

Saliency and potential The (bottom-up) saliency map is derived via the spectral residual (SR) method described in [23] An example is provided in Fig. 4b. We initially experimented with standard saliency from conspicuity maps [26], Bayesian surprise [25] , graph-based visual saliency [20] and self-resemblance [40] methods. However, the SR method provides comparable performance to other methods but at a lower computational complexity end it is easy to code (basically, five Matlab lines [23]).

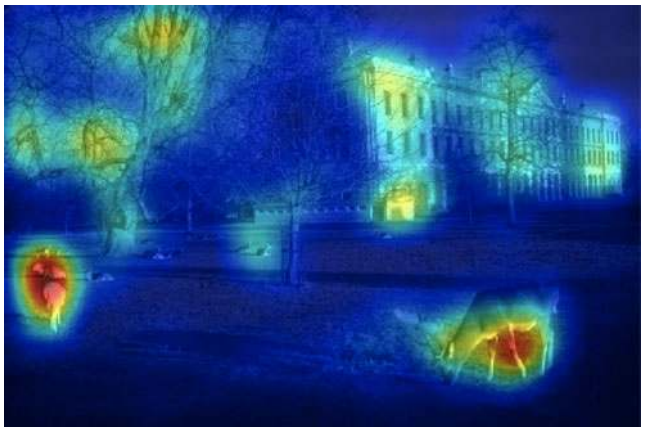
The map is then normalized within the [0, 100] range. From $s(\cdot, t)$, landscape potential $V(\cdot, t)$ is computed via Eq. 4, with $\tau_V = 0.01$ [6]; then, $\nabla V = [\partial V_x, \partial V_y]^T$ is obtained using a finite difference method based on a central difference scheme. The potential surface computed from the saliency map shown in Fig. 4b is presented in Fig. 5a.

Algorithm 1 IHCS Algorithm

Input: Normalized salience map s, number of fixations N_{fix}, Parameters: $\rho, \tau_V, \varepsilon, k_d, \mu, K, \{\alpha, \beta, \gamma, \delta\}, \beta_P, T$
Output: Sequence $\mathbf{r}(1), \mathbf{r}(2), \cdots$ of gaze positions

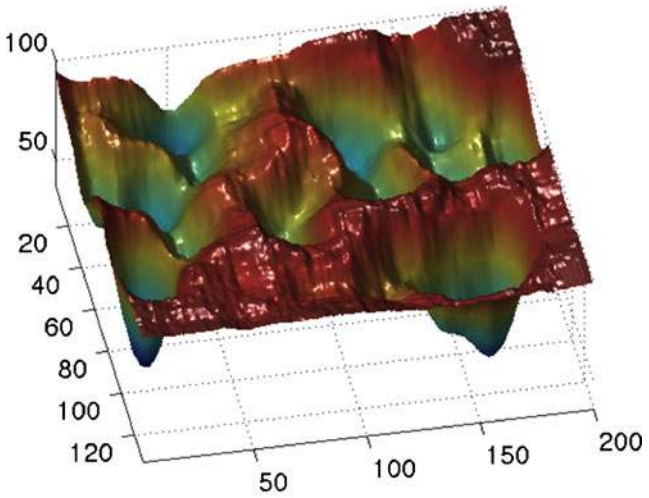
```
Compute potential V, Eq. 4
t ← 1        // gaze-shift counter
n ← 1        // fixation counter
Shift gaze r(t) → center of s
repeat
        // Local search and feeding.....
    Compute the feeding patch N_{r(t)}
    Sample the attraction site set A(t), Eq. 25
    N_S ← |A(t)|
    if N_S > 0 then
        n_s ← 1, ξ ← 1
        while ξ = 1 & n_s ⩽ N_S do
            Shift gaze r(t) → r_new(t), Eq. 14
            Deplete site in N_ε(r_new(t)), Eq. 15
            Update potential, Eq. 4
            Set r(t + 1) ← r_new(t)
            t ← t + 1, n_s ← n_s + 1
                // Action choice
            Compute q, Eq. 12
            Sample ξ ~ Bern(ξ; q)
        end while
        n ← n + 1
    end if
        // Flying.....
        // Motor simulation of Lévy flights
    for k ← 1, K do
        Sample r_{new,k}(t), Eqs. 22, 23, 24
    end for
    Weight the samples through the likelihood Eq. 20.
    Estimate (θ*, l*), Eq. 21
    Shift gaze r(t) → r_new(t), Eq. 6
            // Metropolis step
    Compute Δŝ = ŝ(r_new(t)) − ŝ(r(t))
    if Δŝ > 0 then
        Set r(t) ← r_new(t)
        t ← t + 1, n ← n + 1
    else
        Generate a random number ι
        if ι < exp(Δŝ/T) then
            Set r(t + 1) ← r_new(t)
            t ← t + 1, n ← n + 1
        end if
    end if
until n ≤ N_fix
```

(a) Original image

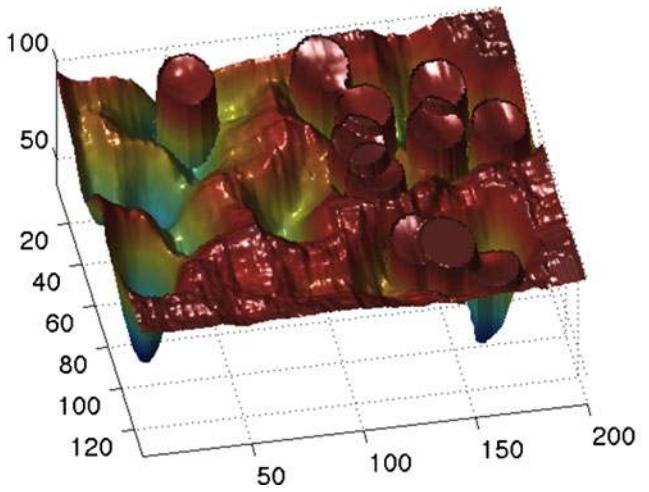


(b) Saliency Map

Fig. 4 The original image (image 23 from the CVPR07 dataset) and the corresponding saliency map obtained via the SR method [23], superimposed on the original image



(a) Potential surface



(b) Potential surface after feeding

Fig. 5 The original potential obtained from the map in Fig. 4b at the beginning of the scanning process and after feeding (10 fixations of IHCS)

Local search and feeding The direct vision range ρ, namely, the radius of the circle $\mathcal{N}_{\mathbf{r}(t)}$, Eq. 3, is set equal to the dimension of the FOA, $|FOA|$.

With regard to IHCS, to implement Eq. 14, a set $\mathcal{A}(t)$ of attraction sites is obtained within the FOA patch by simple thresholding [23],

$$\mathcal{A}(t) = \{\mathbf{r}^*(t) | \mathbf{r}^*(t) \in \mathcal{N}_{\mathbf{r}(t)} \wedge s(\mathbf{r}^*(t)) > 3E[s(\cdot, t)]\}, \quad (25)$$

where $E[s(\cdot, t)]$ is the mean value of s.

Note that for what concerns the HCS method, $\mathbf{r}^*(t)$ is computed within the same range via Eq. 3. This a limit case of Eq. 14 when only the max saliency point is retained.

In IHCS, when the site $\mathbf{r}_{\text{new}}(t)$ is reached, the number n_s of visited points is incremented and destructive foraging is performed in a small neighborhood $\mathcal{N}_\varepsilon(\mathbf{r}_{\text{new}}(t))$ via Eq. 15.

In the simulations presented below, the depletion constant k_d has been set experimentally to 0.2.

For what concerns the radius ε of the depletion region, we assume that such region must cover at least the minimum region covered by a fixation (0.5°, [22]), thus we set, conservatively, a visual angle of $\varphi = 0.6° = 2\varepsilon$. By using as a baseline the same viewing conditions adopted to record the eye-tracking data comprised in the Fixations in FAces dataset [10] (viewing distance $v_{\text{d}} = 80$ cm, screen resolution $s_{\text{r}} = 66.5$ dpi), the diameter d_{fix} of the region $\mathcal{N}_\varepsilon$ can be calculated in pixel units as

$$d_{fix} = \varphi \frac{1}{2\tan^{-1}\left(\frac{1}{2v_{\text{d}}}\right)} \frac{\pi}{180} \frac{s_{\text{r}}}{2.54} \quad (pxl). \quad (26)$$

Thus, $\varepsilon = d_{fix}/2 \approx 10$ pixels.

The effect of feeding is visualized in terms of the potential surface V in Fig. 5b.

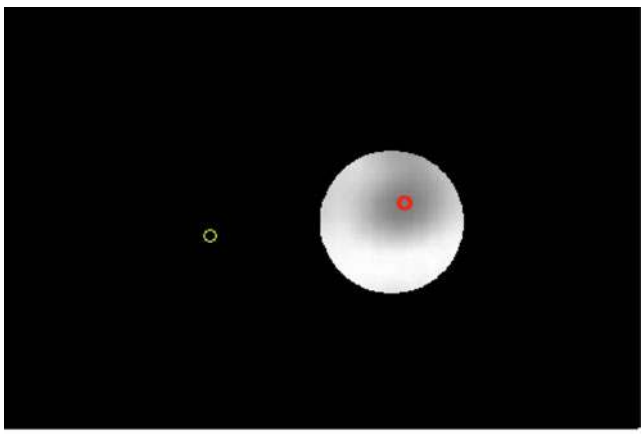
Action choice In IHCS, the choice of the foraging action is sampled from Eq. 16, with q obtained from Eq. 12 where the mean feeding rate μ is experimentally set to 80.

For HCS, the difference of saliency $\Delta s = s(\mathbf{r}^*(t)) - s(\mathbf{r}(t))$ is evaluated and compared with threshold ν so to set the switching variable ξ in Eq. 5; following the ROC-based procedure used for IHCS parameter tuning, ν has been experimentally determined as $\nu = 0.3 \max\{s(\cdot, t)\}$.

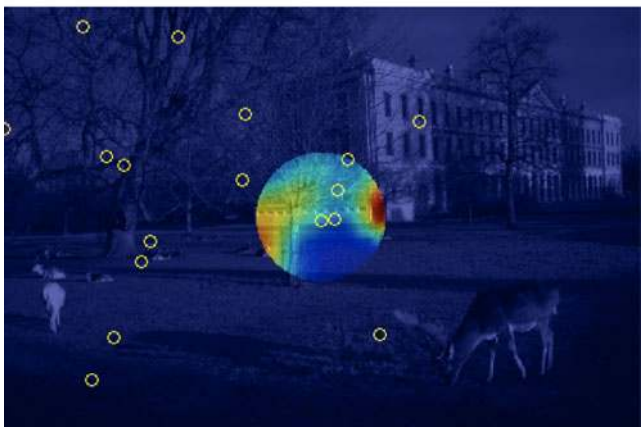
Extensive search via Lévy flights In IHCS, the optimal l, θ components to be chosen according to the MAP rule, Eq. 17, are obtained in practice by the forward sampling procedure articulated in the sampling steps of Eqs. 22–24. In order to ensure a partial visibility condition, the number of generated samples (that is, the candidate gaze shift locations) is set equal to $K = \lfloor \frac{1}{10} \max\{\text{width}, \text{heigth}\} \rfloor$, where width and height are the dimensions of the original image and $\lfloor x \rfloor = \max\{m \in \mathbb{Z} | m \leq x\}$ is the floor operator. The neutral value of $\beta_P = 1$ is used in Eq. 9.

The actual values of the "motor parameters" $\{\alpha, \beta, \gamma, \delta\}$ to be used in the sampling step of Eq. 8 have been derived from the small subset of 10 images randomly chosen from the Fixations in FAces dataset. Given the empirical distributions of gaze shifts, it is possible to fit such distributions in order to derive the parameters of the exhibited α-stable distribution. The estimation of the α-stable distribution is complicated by the aforementioned nonexistence of a closed form pdf. As a consequence, a number of different approximations for evaluating the density have been proposed, see, e.g., [29, 34]. Based on these approximations, parameter estimation is facilitated using the estimator proposed in [29]. Simulation results presented here have been obtained using $\alpha = 1.3, \beta_3 = 1, \gamma = 40, \delta = 0$, where we have set $\delta = 0$, since the drift is accounted for by the deterministic component of Eq. 6

Having fixed the parameters of the α-stable distribution, an α-stable random variable l_k can be sampled - Eq. 24, in several ways. The one applied here is the well known Chambers, Mallows, and Stuck procedure [12].



(a) HCS sampling



(b) IHCS sampling

Fig. 6 Difference between HCS (**a**) and IHCS (**b**) in sampling motor parameters. HCS either chooses the most salient point within direct vision range (*red spot*) or blindly samples a Lévy flight (*yellow circle*); IHCS either samples points within the direct vision range or simulates a set of possible Lévy flights (*yellow circles*) to perform an informed jump

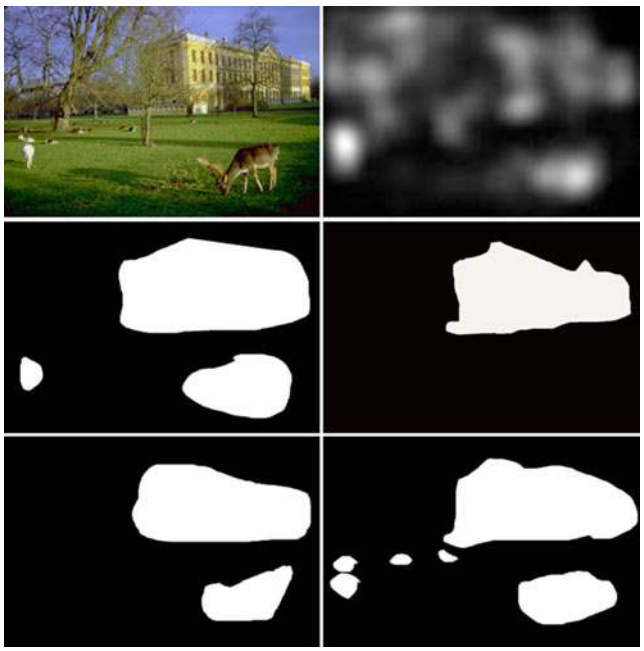


Fig. 7 *Top row*: Image 23 from the CVPR07 dataset and the corresponding saliency map computed as in [23]. *Below*: Segmentation of main objects by four human subjects

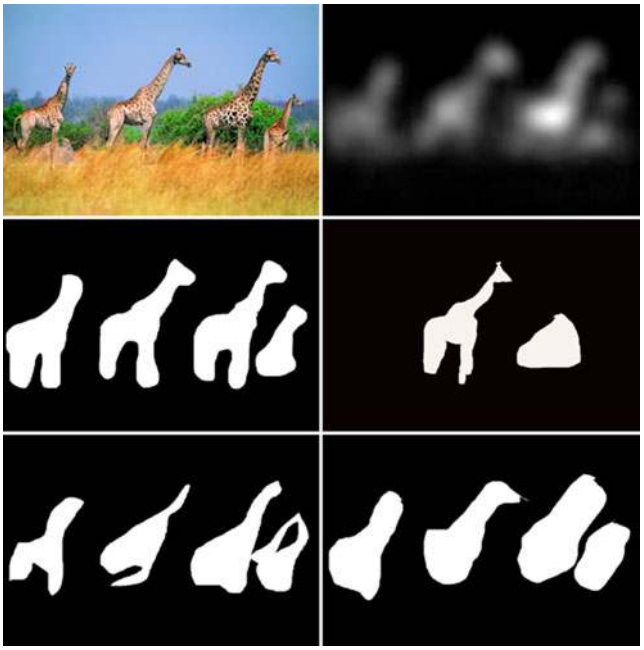


Fig. 8 *Top row*: Image 41 from CVPR07 dataset and the corresponding saliency map computed as in [23]. *Below*: Segmentation of main objects by four human subjects

The decision to accept a candidate flight is accepted according to the Metropolis rule given in Eq. 10. For simulations presented here, $T = 25$ (cfr., [5, 6] for an extensive discussion).

With respect to the determination of a candidate long saccade (Lévy flight), the HCS method is an extreme case of IHCS, in that sampling formalized in Eqs. 7 and 8 can be seen as sampling via Eqs. 22 and 23 using $K = 1$ (that is, only one candidate flight is conjectured). The fundamental difference between IHCS and HCS at this stage is illustrated in Fig. 6.

Finally, after Metropolis evaluation, if no candidate FOA $\mathbf{r}(t)_{\text{new}}$ has been accepted, the current fixation point $\mathbf{r}(t)$ is kept.

The procedure described above is currently implemented in plain MATLAB code, with no specific optimizations and running on a 2.8-GHz Intel Core 2 Duo processor, 2-GB RAM, under Mac OS X 10.6.8. With regard to actual performance of the IHCS under such setting, the average elapsed time for the whole processing amounts to 22.508 s for a 534×800 pixel image. More precisely, 0.36 s is taken to compute saliency,

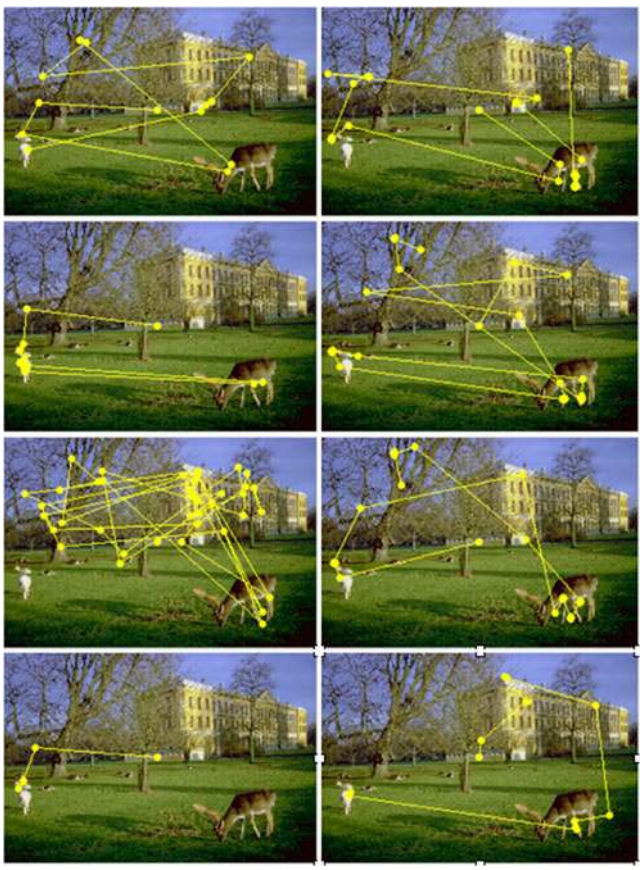


Fig. 9 *Left column*, HCS-generated scanpaths; *right column*, IHCS-generated scanpaths

Fig. 10 *Left column*, HCS-generated scanpaths; *right column*, IHCS-generated scanpaths

Table 1 HR and FAR for the HCS and the IHCS models

Method	HR	FAR
HCS	0.395 ± 0.143	0.109 ± 0.036
IHCS	0.468 ± 0.050	0.104 ± 0.009

while the average elapsed time to sample a scanpath composed by 10 fixations is 22.148 s.

In the same conditions, the HCS method takes an average elapsed time of 32.921 s for the generation of the scanpath. The higher scanpath sampling time is to be related to a low acceptance rate of candidate Lévy flights, due to the uninformed procedure for generating the flight paramters θ, l.

4.3 Experiment 1

The aim of the experiment was to compare IHCS against HCS by quantitatively assessing the exploration performance of the scanpath, given a bottom-up saliency map, with respect to main objects or proto-objects present in the scene. To this end, we exploited the Hou and ZHang CVPR07 dataset. Two examples of the kind of input data are provided in Figs. 7 and 8, showing the original images, the related saliency maps, and the proto-object hand-labeled maps.

The dataset has been processed in order to produce 25 scanpaths for each image by the HCS model with an average of 10 to 15 fixations per scanpath. The 10 images used in the preliminary trial for parameter setting have been discarded. The same was performed by using the IHCS model. Some examples of typical scanpath obtained are shown in Figs. 9 and 10.

It is apparent, from a qualitative standpoint, that IHCS scanpaths have a more plausible behavior with respect to those generated by HCS. In some cases, the HCS might fall in a "potential trap," where no Lévy flights are accepted but the local search is not able to

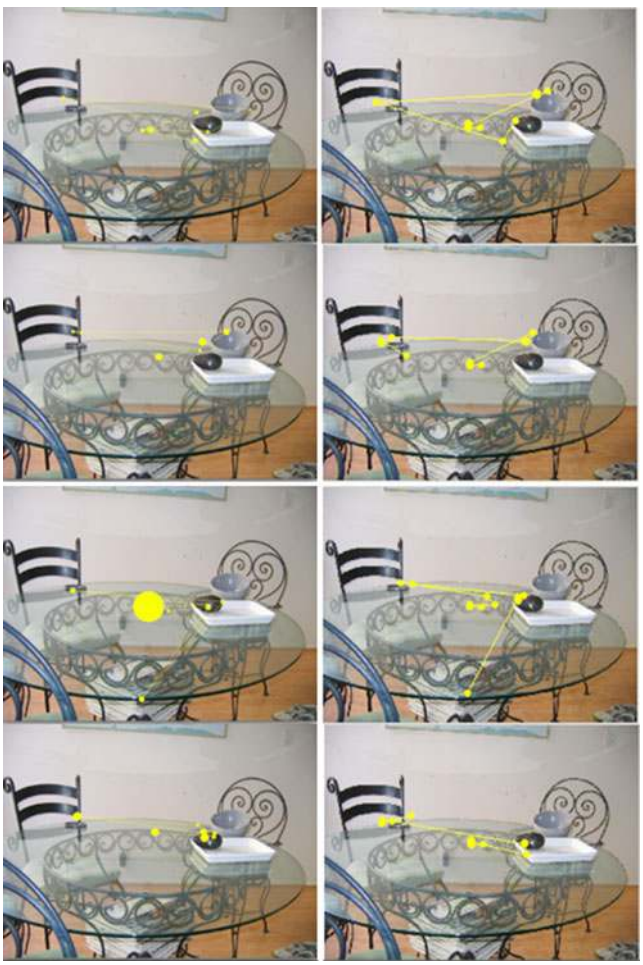


Fig. 11 *Left*, eye-tracked human observers; *right*, IHCS model output

Fig. 12 *Left*, eye-tracked human observers; *right*, IHCS model output

disengage from the fixated site (see, for example, the result in Fig. 9, left image on the bottom row).

Beyond qualitative evaluation, following [23], performance was assessed by measuring the hit rate (HR) and the false alarm rate (FAR) for each image,

$$\mathrm{HR}_s = E\left[\prod_k \mathscr{O}_k \cdot \mathscr{FM}_s\right] \tag{27}$$

$$\mathrm{FAR}_s = E\left[\prod_k (1-\mathscr{O}_k) \cdot \mathscr{FM}_s\right], \tag{28}$$

where $E[]$ denotes expectation and $s = 1, \cdots, 25$ indexes the sth scanpath; $\mathscr{O}_k$ denotes the binary map ($\mathscr{O}_k(x, y) = 1$ for points of target objects, 0 for points in the background) obtained from kth hand labeler; and $\mathscr{FM}_f$ is the binary fixation map obtained by setting to 1 points of the circular region around a fixation of area equal to $1/2|FOA|$ and to 0 points outside such regions. The reason for considering a small foveal region rather than simply the fixation point itself is to provide a different weight for fixations falling in the neighborhood of objects border with respect to fixations occurring within object. Then HR_s and FAR_s are averaged with respect to all scanpaths and to all images. The final total average HR and FAR rates are reported in Table 1, where the better performance of IHCS can be appreciated. It is worth noting that on single images with many objects to be visited (and coherent object maps between subjects), such difference is much higher (for instance in the case of the Giraffe image, the IHCS HR is approximately 0.6 and the HCS HR is 0.3). It is worth noting that, with regard to computational efficiency, the IHCS runs in a number of iterations of the feed and fly cycles, roughly corresponding to the number of fixations desired; in order to obtain a comparable number of fixations (10–15) via HCS, a much higher number of iterations must be exploited (here

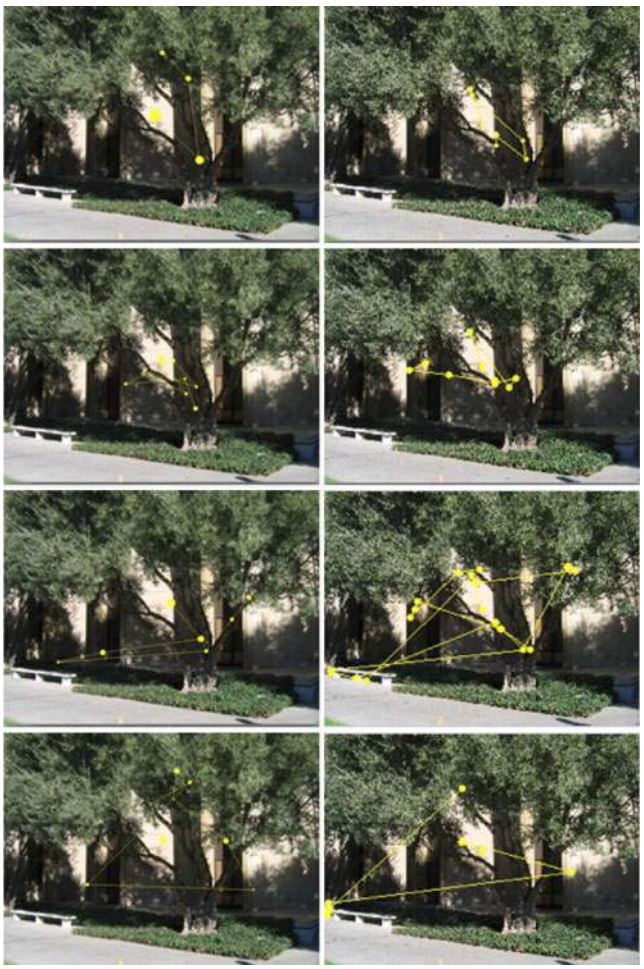


Fig. 13 *Left*, eye-tracked human observers; *right*, IHCS model output

Fig. 14 *Left*, eye-tracked human observers; *right*, IHCS model output

150) due to the higher rejection rate of the proposed Lévy flights than in the IHCS run.

4.4 Experiment 2

The aim of this experiment was to qualitatively compare the motor behavior represented in the IHCS-generated scanpaths with scanpaths eye tracked from human subjects. For this experiment, we used the Fixations in FAces dataset. We generated 20 scanpaths for each image and compared them to those exhibited by human observers by choosing most similar scanpaths in terms of fixations coordinates, duration, and time occurrence. Some typical results are shown in Figs. 11, 12, and 13 showing the ability of IHCS to mimic observer oculomotor behavior ($K = 100$).

Further, we explored the condition in which the same images contained ecologically relevant objects like faces and tested the capability of IHCS to provide plausible scanpaths. To this end, we combined the bottom-up saliency map with the detected face map, in the vein of [10, 11]. The face saliency map is formed by convolving delta functions at the detected facial centers with 2-D Gaussians having standard deviation equal to the estimated facial radius [11]. The values of this map were normalized to a fixed range and linearly added to the bottom-up saliency map, so to obtain a master saliency map. Here, the number of fixations per image was 6/7; the other critical parameters of the method (μ, k_d, T) were the same as in the no face experiment. The scanpaths produced are compared to human eye-tracked data in Figs. 14, 15, and 16.

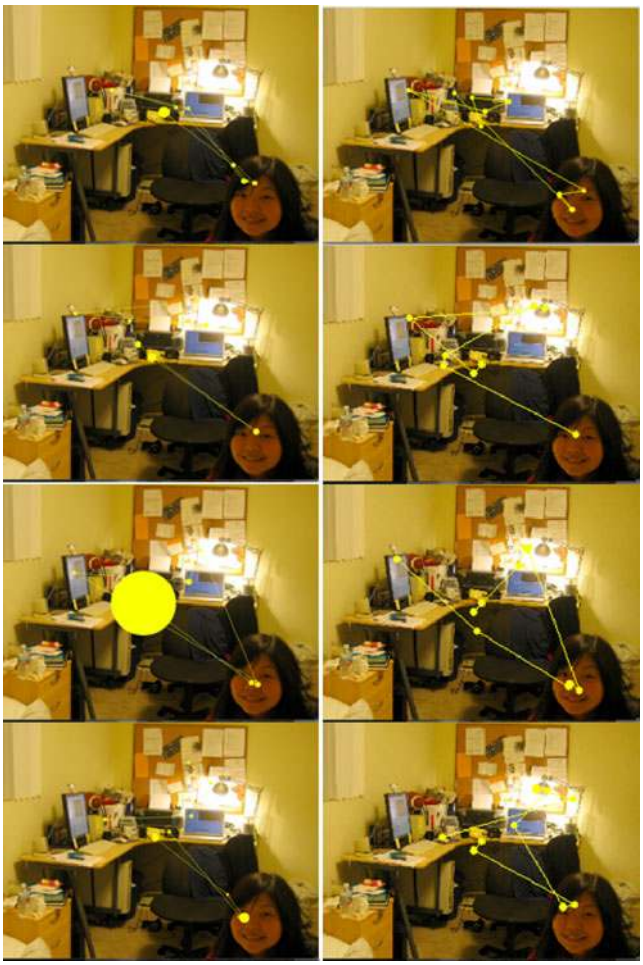


Fig. 15 *Left*, eye-tracked human observers; *right*, IHCS model output

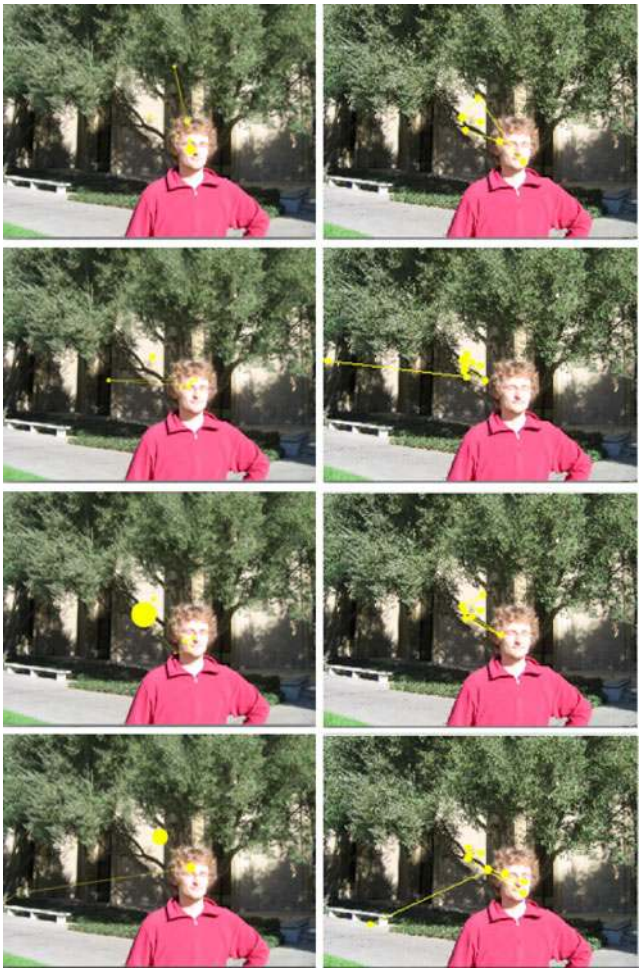


Fig. 16 *Left*, eye-tracked human observers; *right*, IHCS model output

5 Conclusions

The IHCS model for the generation and control of scanpaths has been presented. The model accounts

for the randomness of visual exploration exhibited by different observers when viewing the same scene, or even by the same subject along different trials. This is a point that is neglected by the great majority of foveation models, although it could be critical for applications of foveation image processing like image/video coding [7, 24, 32, 46], image/video retrieval [15], and quality assessment [47], but it is also relevant to computer vision and learning tasks.

The rationale behind the work presented here is that the exploitation of systematic tendencies characterizing oculomotor behaviors [43] can be an advantage for simulating the visual sampling of the surrounding world. More generally, this approach may be developed for a principled modeling of individual differences, a key issue in cognitive science [44], since providing cues for defining the informal notion of scanpath idiosyncrasy in terms of individual gaze-shift distribution parameters.

The model, which further extends and improves on the HCS model presented in [6], generates sequences of fixations and gaze shifts under the control of an information foraging mechanism implemented through a stochastic dynamical system that switches between two states: "feed" and "fly." A novelty introduced in the present study is an internal gaze-shift simulation step to estimate the best motor parameters for the actual shift, akin to visibility or value-based models of eye movement behaviors [30, 39].

The simulations show that the method is independent from features adopted to derive saliency and can reliably cope with either bottom-up or top-down semantic cues. Also, the method could be easily extended to embed object-based paradigms. For instance, rather than looking for a point with large saliency values the model could be amended to give priority to fixations dwelling upon regions/patches representing objects or proto-objects that have relevance in determining organism behavioral responses [8, 18].

References

1. van Beers R (2007) The sources of variability in saccadic eye movements. J Neurosci 27(33):8757–8770
2. Begum M, Karray F (2011) Visual attention for robotic cognition: a survey. IEEE Trans Auton Mental Develop 3(1): 92–105
3. Bishop CM (2006) Pattern recognition and machine learning (Information Science and Statistics). Springer, New York
4. Boccignone G, Chianese A, Moscato V, Picariello A (2005) Foveated shot detection for video segmentation. IEEE Trans Circuits Syst Video Technol 15(3):365–377
5. Boccignone G, Ferraro M (2004) Modelling gaze shift as a constrained random walk. Physica A: Statistical Mechanics and its Applications 331(1–2):207–218
6. Boccignone G, Ferraro M (2011) Modelling eye-movement control via a constrained search approach. In: Proceedings of 3rd European workshop on visual information processing (EUVIP 2011). IEEE Press, Piscataway, pp 235–240
7. Boccignone G, Marcelli A, Napoletano P, Di Fiore G, Iacovoni G, Morsa S (2008) Bayesian integration of face and low-level cues for foveated video coding. IEEE Trans Circuits Syst Video Technol 18(12):1727–1740
8. Borji A, Itti L (2012) State-of-the-art in visual attention modeling. IEEE Trans Pattern Anal Mach Intell. http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.89
9. Brockmann D, Geisel T (2000) The ecology of gaze shifts. Neurocomputing 32(1):643–650
10. Cerf M, Frady E, Koch C (2009) Faces and text attract gaze independent of the task: experimental data and computer model. J Vis 9(12):10.1–10.15
11. Cerf M, Harel J, Einhäuser W, Koch C (2008) Predicting human gaze using low-level saliency combined with face detection. In: Advances in neural information processing systems, vol 20. MIT Press, Cambridge, pp 545–552
12. Chambers J, Mallows C, Stuck B (1976) A method for simulating stable random variables. J Am Stat Assoc 71(354): 340–344
13. Churchland P, Ramachandran V, Sejnowski T (1994) A critique of pure vision. MIT Press, Cambridge
14. Codling E, Plank M, Benhamou S (2008) Random walk models in biology. J R Soc Interface 5(25):813
15. Cotsaces C, Nikolaidis N, Pitas I (2006) Video shot detection and condensed representation. a review. IEEE Signal Process Mag 23(2):28–37
16. Da Luz M, Buldyrev S, Havlin S, Raposo E, Stanley H, Viswanathan G (2001) Improvements in the statistical approach to random Lévy flight searches. Physica A: Statistical Mechanics and its Applications 295(1–2):89–92
17. Ellis S, Stark L (1986) Statistical dependency in visual scanning. Hum Factors 28(4):421–438
18. Frintrop S, Rome E, Christensen H (2010) Computational visual attention systems and their cognitive foundations: a survey. ACM Trans Appl Percept 7(1):1–39
19. Gnedenko B, Kolmogórov A (1954) Limit distributions for sums of independent random variables. Addison-Wesley, Reading
20. Harel J, Koch C, Perona P (2007) Graph-based visual saliency. In: Advances in neural information processing systems, vol 19. MIT Press, Cambridge, pp 545–552
21. Harris C (1998) On the optimal control of behaviour: a stochastic perspective. J Neurosci Methods 83(1):73–88
22. Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Van de Weijer J (2011) Eye tracking: a comprehensive guide to methods and measures. Oxford University Press, Oxford
23. Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. In: Proceedings CVPR 07, vol 1, pp 1–8
24. Itti L (2004) Automatic foveation for video compression using a neurobiological model of visual attention. IEEE Trans Image Process 13(10):1304–1318
25. Itti L, Baldi P (2009) Bayesian surprise attracts human attention. Vis Res 49(10):1295–1306
26. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20:1254–1259
27. Jackson, J (1958) Evolution and dissolution of the nervous system (Croonian lectures). Published in parts in the British Medical Journal, Lancet pp 5–75
28. Klein R, MacInnes W (1999) Inhibition of return is a foraging facilitator in visual search. Psychol Sci 10(4):346–352

29. Koutrouvelis I (1980) Regression-type estimation of the parameters of stable laws. J Am Stat Assoc 75(372):918–928
30. Kowler E (2011) Eye movements: the past 25 years. Vis Res 51(13):1457–1483. 50th Anniversary Special Issue of Vision Research - vol 2
31. Kustov A, Robinson D (1996) Shared neural control of attentional shifts and eye movements. Nature 384:74–77
32. Lee J, De Simone F, Ebrahimi T (2011) Efficient video coding based on audio-visual focus of attention. J Vis Commun Image Represent 22(8):704–711
33. Neal R (1993) Probabilistic inference using Markov chain Monte Carlo methods. Department of Computer Science, University of Toronto. http://www.cs.utoronto.ca/radford/
34. Nolan J (1997) Numerical calculation of stable densities and distribution functions. Commun Stat Stoch Models 13(4):759–774
35. Noton D, Stark L (1971) Scanpaths in eye movements during pattern perception. Science 171(968):308–311
36. Plank M, James A (2008) Optimal foraging: Lévy pattern or process? J R Soc Interface 5(26):1077
37. Privitera CM, Stark LW (2000) Algorithms for defining visual regions-of-interest: comparison with eye fixations. IEEE Trans Pattern Anal Mach Intell 22(9):970–982
38. Reynolds A (2008) Optimal random Lévy-loop searching: new insights into the searching behaviours of central-place foragers. EPL (Europhysics Letters) 82(2):20001.1–20001.6
39. Schütz A, Braun D, Gegenfurtner K (2011) Eye movements and perception: a selective review. J Vis 11(5):9
40. Seo H, Milanfar P (2009) Static and space-time visual saliency detection by self-resemblance. J Vis 9(12):1–27
41. Stark L, Privitera C, Yang H, Azzariti M, Ho Y, Blackmon T, Chernyak D (2001) Representation of human vision in the brain: how does human perception recognize images? J Electron Imaging 10:123–151
42. Stephen D, Mirman D, Magnuson J, Dixon J (2009) Lévy-like diffusion in eye movements during spoken-language comprehension. Phys Rev E 79(5):056114.1–056114.6
43. Tatler B, Vincent B (2009) The prominence of behavioural biases in eye guidance. Vis Cogn 17(6–7):1029–1054
44. Vandekerckhove J, Tuerlinckx F, Lee M (2011) Hierarchical diffusion models for two-choice response times. Psychol Methods 16(1):44
45. Viswanathan G, Afanasyev V, Buldyrev S, Havlin S, Da Luz M, Raposo E, Stanley H (2000) Lévy flights in random searches. Physica A: Statistical Mechanics and its Applications 282(1–2):1–12
46. Wang Z, Lu L, Bovik AC (2003) Foveation scalable video coding with automatic fixation selection. IEEE Trans Image Process 12:1–12
47. You J, Reiter U, Hannuksela M, Gabbouj M, Perkis A (2010) Perceptual-based quality assessment for audio–visual services: a survey. Signal Process Image Commun 25(7):482–501