Gaze shift behavior on video as composite information foraging

Signal Processing: *Image Communication* 28 (2013) 949–966

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Gaze shift behavior on video as composite information foraging

G. Boccignone [a],*, M. Ferraro [b]

[a] *Dipartimento di Informatica, Università di Milano, via Comelico 39/41, Milano 20135, Italy*
[b] *Dipartimento di Fisica, Università di Torino, via Giuria 1, 10125 Torino, Italy*

ARTICLE INFO

Available online 21 July 2012

Keywords:
Visual attention
Eye movements
Random walk
Active vision
Information encoding

ABSTRACT

The ability to predict, given an image or a video, where a human might fixate elements of a viewed scene has long been of interest in the vision community. However, one point that is not addressed by the great majority of computational models is the variability exhibited by different observers when viewing the same scene, or even by the same subject along different trials. Here we present a model of gaze shift behavior which is driven by a composite foraging strategy operating over a time varying visual landscape and accounts for such variability.

The system performs a deterministic walk if in a neighborhood of the current position of the gaze there exists a point of sufficiently high saliency; otherwise the search is driven by a Langevin equation whose random term is generated by an α-stable distribution.

Results of the simulations on complex videos from the publicly available University of South California CRCNS eye-1 dataset are compared with eye-tracking data and show that the model yields gaze shift motor behaviors that exhibits statistics similar to those exhibited by human observers.

1. Introduction

Visual attention guides our gaze to relevant parts of the viewed scene. At the same time, the urge for moving the eyes lies with the need for bringing the fovea to sample information not fully available in the limited acuity peripheral vision.

Thus, the aim of a computational model of visual attention is to answer the question *Where to Look Next*? by providing, at the computational theory level [1], an explanation of the mapping *viewed scene* $\mapsto$ *gaze sequence*, together with a procedure (the algorithmic level, [1]) that implements such mapping.

Modeling visual attention has been a very active research area over the past 25 years (see [2,3] and, more specifically, [4–6] for computational approaches). Yet, we are still some way from a model that can explain many of the complexities of eye movement behavior.

One of the most intriguing conundrums is the variability of eye movements. When looking, for instance, at natural movies under a free-viewing or a general-purpose task (e.g., "follow main actors and actions"), the moment-to-moment relocation of gaze can be different among observers even though the same locations are taken into account. Variability is exhibited by the same subject along different trials on equal stimuli. Further, the consistency in fixation locations between observers decreases with prolonged viewing [7].

An example of such behavior is provided in Fig. 1.

This effect is even more noticeable when free-viewing static images: consistency in fixation locations selected by observers decreases over the course of the first few fixations after stimulus onset [8] and can become idiosyncratic. Only in natural settings a higher degree of spatial and temporal consistency can be found [9], when eye

* Corresponding author.
E-mail addresses: Giuseppe.Boccignone@unimi.it (G. Boccignone), ferraro@ph.unito.it (M. Ferraro).
URL: http://boccignone.di.unimi.it/ (G. Boccignone).

0923-5965/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.image.2012.07.002

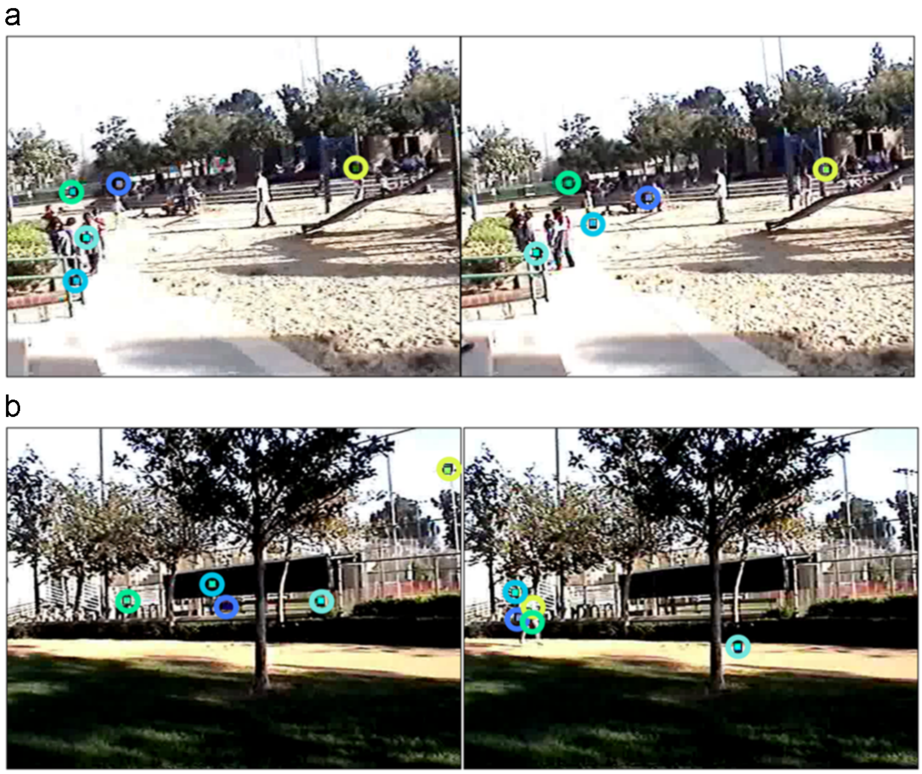

Fig. 1. Gaze positions (centers of the colored circles) recorded from five different observers on two frames of the `beverly01` clip (a) and from the `beverly08` clip (b) both from the CRCNS data set.

movements are involved in gathering information to actually accomplish behavioral tasks (driving, walking, etc.).

Although there is a small probability that two observers will fixate exactly the same location at exactly the same time, nevertheless, eye movements of several observers on the same natural movie are less variable than on different movies. In addition, a higher degree of inter-observer coherence has been reported when isolated objects start to move [7], witnessing the fact that eye movements are at least partially determined by the visual input (one example is provided in the right frame of Fig. 1(b)).

To summarize, where we choose to look next at any given moment in time is not completely deterministic, yet neither is it completely random [10]. Under these circumstances, the question with which we are concerned can be put as follows: is a computational model of gaze shift conceivable that accounts for and mimics such interplay between determinism and randomness? So far, the question has been surprisingly overlooked by most computational models of visual attention (cf., [4]). Nevertheless, the problem highlighted, beyond its theoretical relevance (variability as intrinsic to optimal control, rather than being simply "noise" [11]), is crucial for behavioral learning (e.g., in robotics [12,13]) and could be beneficial to many fields of applications of visual attention modelling [4].

We propose in the present paper a stochastic model of eye guidance on natural videos, where the eye's behavior is that of an information foraging random walker. The walker samples the time-varying visual landscape driven by a composite mechanism akin to ecological models of animal foraging: an *extensive stage* provides global relocations of gaze, shaped in the form of Lévy flights, to new regions of the landscape; an *intensive stage* allows for local scanning and information gathering.

An early conjecture of such mechanism, but limited to handling static images, has been discussed in [14]. Here important novelties are introduced. First, when videos are considered, the intensive stage can account for both fixations and smooth pursuit. Second, a richer set of Lévy flights is exploited to perform the extensive search by sampling the gaze shift amplitudes from the general family of α-stable distributions [15], rather than being committed to the Cauchy distribution as in [16,17,14]; the parameters governing such distributions can be inferred from eye-tracking data thus providing a higher degree of agreement between model-generated and natural oculomotor behaviors. Third, and most important, an "internal simulation" step is performed in order to infer the posterior distribution of shift amplitudes and directions, which is then exploited to determine the actual relocation of the foraging eye.

In the following section we provide some crucial background to lay down the rationales and the assumptions of the model. In Section 3 we formalize such assumptions together with the mechanisms to perform intensive and extensive behavior. In Section 4 we provide simulations on complex videos, from a publicly available dataset; we also compare with eye-tracking data to assess whether the gaze shift motor behaviors generated by the model are characterized by statistics related to behaviors generated by human observers. Finally, in Section 5, we discuss results so far achieved, while relating them to current trends in the visual attention literature, and address some limitations of the current version of the model.

2. Background

We briefly review some pivotal issues related to the problem of gaze shift variability that are central for setting out our approach.

Any computational model of visual attention has to face four major constraints on the mapping *viewed scene* $\mapsto$ *gaze sequence* : (1) the form of the input, (2) the task assigned to the observer; (3) the levels of control of eye guidance; (4) the sensory and motor noise. These are intertwined with one another.

Input. The input can be provided in the form of either a picture (static image) $\mathbf{I}$, or a video, that is a time-parametrized sequence of images $\{\mathbf{I}(1),\mathbf{I}(2),\ldots,\mathbf{I}(t),\ldots\}$, or a real-world setting where the situated observer, either natural or artificial, is engaged in some behavioral task. Pictures can be conceived as sudden whole-scene onsets that have no real-world analogue; further, they lack spatio-temporal information. For these reasons pictures fall short of ecological plausibility and are poor surrogates for real environments [9]. Videos include dynamic information and constitute a form of input stimuli of higher ecological value with respect to static images; this is true for natural videos, whilst some caution should be exercised with "Hollywood-style" or "MTV-style" movies where editorial cuts may result in oculomotor disruptions to normal scene perception [9,7]; again, cuts can give rise to unnatural abrupt whole-scene onsets.

Whatever the form of the input, many visual attention models do not account for the retinal position of image information, and decreasing retinal acuity in the periphery is overlooked [9]. Yet, retinal anisotropies in sampling play a role in tendencies to move the eyes in particular ways, and the assumption of uniform spatial sampling can lead to distributions of saccade amplitudes that do not match human eye behavior [9].

Task. A distinct gaze sequence pattern is associated with a specific imposed task [18,19]. For instance, when dealing with pictures the classes of tasks that can be given are only a subset of the repertoire of behaviors we execute in the real world (a further reason to consider pictures inadequate proxies for real environments [9]). In the real world most fixations are on task-relevant objects [10].

Interestingly, it has been shown that tasks can be characterized by low-level eye movement metrics such as fixation duration and gaze shift amplitudes, and by the shape of their distributions: for instance, walking and talking show broad distributions; reading, counting, and sorting show narrower distributions [10]. What is worth recalling here is that free-viewing as a task-free condition should be handled with extreme caution. In recent research, criticism has been raised against such setting, for it entitles the observer to arbitrarily select his own internal agendas [9].

Level of control. In analogy with other aspects of motor behavior and action selection [20–22], the guidance of eye movements is likely to be influenced by a hierarchy of several interacting control loops, operating at different levels of processing of the whole action-perception loop. Each processing level exploits the most suitable representation $\mathcal{R}$ of the viewed scene for its own level of abstraction: Schütz et al. [2], in a plausible portrayal, have singled out salience, objects, values, and plans.

There is a large literature on models that have been conceived in order to account for either one specific or multiple joint levels of abstraction, which we will not review here (see, in particular, [2,4]). But we wish to draw attention to two issues. First, up to this date, the majority of computational models has retained a central place for low-level visual conspicuity or salience; for instance, the most recent review by Borji and Itti [4] aims at surveying models that can compute saliency maps, but as a matter of fact it covers most of the models currently adopted in many applications. Yet, the explanatory power of saliency has been strongly questioned (cf., [9] for a deep and lucid account). Indeed, prediction of fixation locations does not imply a true causal influence, it might be the case that salience merely covaries with another factor, which is actually controlling gaze [2].

Second, a great deal of approaches that qualify as computational models of visual attention are incomplete with respect to the mapping *viewed scene* $\mapsto$ *gaze sequence*. As a matter of fact, these approaches concentrate on computing a mapping from an image, or less frequently from an image sequence, to a representation, typically a saliency map s. The saliency map is then quantitatively evaluated by comparing with eye movement data according to some evaluation measure [4]. Thus, a partial mapping $\mathbf{I}\mapsto s$ is provided in lieu of a mapping to a temporal sequence of gaze positions $\mathbf{I}\mapsto\{\mathbf{r}(1),\mathbf{r}(2),\ldots\}$.

Clearly, even though the mapping $\mathbf{I}\mapsto s$ is taken for granted, yet the next step $s\mapsto\{\mathbf{r}(1),\mathbf{r}(2),\ldots\}$ is a long way off. It has been shown that, when viewing static images, observers exhibit oculomotor biases or systematic tendencies [23,24]. These can be thought of as regularities that are common across all instances of and manipulations to the behavior. For instance, such tendencies can be detected in saccade amplitudes: motor biases in the oculomotor system are likely to promote small amplitude saccades rather than large amplitude saccades. Thus, amplitudes show a positively skewed, long-tailed distribution in most experimental settings in which complex scenes are viewed [25,23,24]. Similar tendencies have been reported on natural movies [7].

The elegant experiments performed by Tatler and Vincent [24] have shown that (i) a model based on oculomotor biases alone performs better on pictures than the standard salience model; (ii) when motor tendencies are formalized as prior probabilities on saccade amplitudes and angles, and such priors are used to weight the likelihood of observed data the predictive performance of a saliency-based model can be dramatically improved. Summing up, oculomotor biases are informative and their statistics account for a number of *hidden sources* such as biomechanics/energetical factors, search efficiency and uncertainty reduction, distribution of objects of interest in the environment, task parameters and higher level cognitive factors [24].

Oculomotor biases can also be considered as mechanisms at least tied to general regulative adjustment of ocular behavior applicable to many visual information-processing tasks, if not part of strategies that are optimal to minimize search time and maximize accuracy [26,27].

Noise. Beyond systematic tendencies and strategies, it is worth recalling that sensory and motor noise limits the precision with which we can sense the world and act upon it [28,29]. A part of the variance in measured eye movements parameters – in analogy with limb motor control mechanisms – may be attributed to random fluctuations in the system, in particular to stochastic variability in neuromotor force pulses [30–33]. For instance, under such constraints, most theories of reading eye movements assume that eye-movement control is inherently probabilistic [34].

3. The model

3.1. Rationales and assumptions

The goal of the present study is to develop a model that describes statistical properties of gaze shifts as closely as possible. In order to account for the several factors in the perception/action loops involved in the guidance of eye-movements (see discussion above), we assume that the gaze sequence $\{\mathbf{r}(1),\mathbf{r}(2),\ldots\}$ is generated by an underlying stochastic process and that the basic parameters characterizing such sequence, namely gaze shift magnitudes l and directions θ, are random variables.

The input is in the form of a foveated RGB video, namely a sequence of time parameterized color images $\{\mathbf{I}(1),\mathbf{I}(2),\ldots,\mathbf{I}(t),\ldots\}$, where t is the time parameter. Each color image $\mathbf{I}(t)$ is a vector, a mapping from the support $\Omega \subseteq R^2$ to an m-dimensional range, $\mathbf{I}:\Omega \to C \subseteq R^m$. At any moment t, the actual input is obtained through the mapping $\mathbf{I}(t) \mapsto \widehat{\mathbf{I}}(t)$ where the foveated frame $\widehat{\mathbf{I}}(t)$ is obtained as a function of current gaze position $\mathbf{r}(t) \in \Omega$ (in the sequel, we will drop the "hat" notation for simplicity, and write $\mathbf{I}(t)$ to denote the foveated frame).

The video can have complex content, displaying moving objects such as pedestrians, cars, cyclists distributed across the screen. This somehow constrains and simplifies the task-assignment problem, an issue that will be discussed later on in Section 4.1.

Though the model *per se* does not strictly depend on a particular representation, in this study, due to the kind of the input, the representation $\mathcal{R}$ is assumed to be the spatio-temporal saliency; the mapping $\mathbf{I}(t) \mapsto s(t)$ may be actually computed with any suitable state-of-the-art method [4]. The limitations of using low-level saliency as a representation of the viewed scene [9] are mitigated here due to the content of the movies we are dealing with, displaying one or multiple moving objects. So far local features and objects are often correlated and saliency maps computed on dynamic features are more predictive than those relying on static features [35]. Thus, low-level salience plays a much bigger role for determining the gaze sequence $\{\mathbf{r}(1),\mathbf{r}(2),\ldots\}$, be it a causal or correlational effect [2].

Under such preliminary assumptions, we are then left with the problem of giving form to the mapping $s \mapsto \{\mathbf{r}(1),\mathbf{r}(2),\ldots\}$. This amounts to answering the question: What kind of stochastic process can give rise to the observed eye movement patterns?

From eye-movements studies we see evidence for periodic large amplitude relocations to new regions, followed by small amplitude shifts exploring the new region [10,25,23,24,7,26,9]; this tendency is strikingly similar, with respect to the resulting movement patterns and their statistics to those exhibited by foraging animals [36–39]. In other terms, eye movements and animal foraging address in some way a similar problem. Building on this rationale, in [17] a gaze shift model has been proposed. Such model is akin to models of simple animal foraging, where the "foraging eye" by engaging in Lévy flights, hunts for areas that are rich in visual saliency. Lévy flights, as opposed, for instance, to usual random walk, may be essential for optimal search in foraging [36].

However, the general applicability of Lévy flights in ecology and biological sciences is still open to debate. Recent experimental data show that the movement patterns of various marine predators and terrestrial animal exhibit a Lévy walk pattern, corresponding to an *extensive* search, in areas with low abundance of preys or foods and an *intensive* feeding resulting, for instance, in a local walk pattern (a sort of food tracking) in areas with high abundance [37]. Thus, in complex environments, optimal searches are likely to result from a mixed/composite strategy [38,39].

An early conjecture in the direction of using a composite strategy for programming gaze shift, limited to handling static images, has been discussed in [14], showing higher efficiency and effectiveness, with respect to [17], in visiting salient the areas of the image. Expanding in scope such conjecture, the gaze sequence $\{\mathbf{r}(1),\mathbf{r}(2),\ldots\}$ is generated by a "foraging eye", which is conceived as random walker that hunts for moving or stationary preys by sampling the time-varying visual landscape $\{s(1),s(2),\ldots,s(t),\ldots\}$.

Such walker performs composite information foraging suitable to implement a coarse to fine strategy of visual scanning: during an *extensive stage* it provides global relocations of gaze to new regions of the landscape; along an intensive stage, it performs local scanning and information feeding.

The steps in making the model are (1) define a global relocation of gaze suitable to implement a saccadic shift, (2) define the intensive stage where the walker closely follows (smooth pursuit) or captures (fixation) some "prey" and (3) allow the forager to choose between performing local search and leaving for a global scanning of the landscape. Steps (2) and (3) will be designed similarly to [14], the major difference being that, in this study, during the intensive stage we account for both fixations and smooth pursuit. Step (1) relies on Lévy flights coupled with internal simulation and is described below.

3.2. Extensive relocation through internal simulation

Let $\mathbf{r}(t) \in \Omega$ be the current position of gaze and $\mathbf{r}_{new}(t)$ a possible new position. During the extensive stage, the motion of the walker $\mathbf{r}(t)\rightarrow\mathbf{r}_{new}(t)$ (the gaze shift) performed on the visual landscape can be conceived as a random walk driven by an external potential V and can be formalized in terms of the Langevin-type equation

$$\mathbf{r}_{new}(t)=\mathbf{r}(t)-\nabla V+\boldsymbol{\eta}, \tag{1}$$

where $\mathbf{r}_{new}(t)$ is determined by the gradient of the potential ∇V and by a stochastic vector with components

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

The θ (flight direction) and l (jump length) random variables stochastically summarize the internal action choice of the forager [24].

The drift term $-\nabla V$ represents the external force field which is shaped by the time-varying saliency landscape, since following [17,14], we define the scalar field V as a decreasing function of the saliency s,

$$V(x,y,t)=\exp(-\tau_V s(x,y,t)), \tag{3}$$

where $-\tau_V$ is a suitable damping parameter. This way a site of high saliency is transformed in a potential hole that, in the absence of a random force, deterministically drives the motion toward its minimum.

To gain some insight on the type of motion that can be performed via Eq. (1) consider for simplicity a zero-drift displacement $\mathbf{r}_{new}(t)=\mathbf{r}(t)+\boldsymbol{\eta}$. Under a uniform distribution of possible directions, the motion is determined by the probability density function f from which shift amplitudes l are sampled, $l\sim p(l)$. For instance, if $p(l)$ is a Gaussian distribution, the usual Brownian motion occurs. However, Brownian motion is a rather inefficient foraging strategy and, further, when applied to model saccadic gaze shifts it does not account for the shape of their distribution [16]; thus, Lévy flights have been proposed by specifying $p(l)$ as a Cauchy distribution [17,14].

In the present study, in order to endow the extensive stage with the capability of large amplitude relocations to new regions, followed by small amplitude shifts, we resort to the general family of α-stable distributions [15] (cf., Appendix A). These form a four-parameter family of continuous probability densities, say $f(l;\alpha,\beta,\gamma,\delta)$, 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. These include as particular cases the Gaussian and Cauchy distributions. By assuming $p(l)=f(l;\alpha,\beta,\gamma,\delta)$, the probability density function of length jumps scales, asymptotically, as $l^{-1-\alpha}$, and thus relatively long jumps are more likely when α is small. 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 length jump is "broad" and the so called Lévy flights take place. Clearly, by modulating $\alpha\in[1;2($, a variety of Lévy flight behaviors can be inferred/simulated.

Thus, α-stable distributions provide the most general solution to generate via Eqs. (1) and (2) random walks exhibiting periodic large amplitude relocations to new regions, followed by small amplitude shifts. Such kind of motion gives rise to the positively skewed, long-tailed distribution of amplitudes that have been observed for long time in the eye-movement literature [40,41,25,23,24,9].

Note at this point that if θ and l were straightforwardly sampled from the priors $p(\theta)$ and $p(l)$, respectively, and inserted in (2), a classic Lévy flight driven by external potential would occur, which slightly improves on the methods described in [16,17,14], for the latter were constrained to work with the specific case of the Cauchy distribution ($f(l;1,0,\gamma,\delta)$).

Here, more generally, we assume that the actual parameters (θ^{*},l^{*}) to be used within Eq. (2) are chosen in order to maximize the posterior distribution $p(\theta,l|s(\mathbf{r}_{new}(t)),s(\mathbf{r}(t)),\mathbf{r}_{new}(t),\mathbf{r}(t))$,

$$(\theta^{*},l^{*})=\arg\max_{\theta,l}p(\theta,l|s(\mathbf{r}_{new}(t)),s(\mathbf{r}_{new}(t)),\mathbf{r}_{new}(t),\mathbf{r}(t)). \tag{4}$$

More precisely, the selection of action parameters should be conditioned on the gain achievable by shifting to a new information state $(\mathbf{r}_{new}(t),s(\mathbf{r}_{new}(t)))$ 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 $p(\theta,l,s(\mathbf{r}_{new}(t)),s(\mathbf{r}(t)),\mathbf{r}_{new}(t),\mathbf{r}(t))$. The latter can be factorized as

$$\begin{aligned}&p(\theta,l,s(\mathbf{r}_{new}(t)),s(\mathbf{r}_{(}t)),\mathbf{r}_{new}(t),\mathbf{r}(t))\\&\quad\approx p(s(\mathbf{r}_{new}(t))|s(\mathbf{r}(t)),\mathbf{r}_{new}(t),\mathbf{r}(t))\\&\quad p(\mathbf{r}_{new}(t)|\mathbf{r}(t),\theta,l)p(\theta)p(l).\end{aligned} \tag{5}$$

The first factor provides the likelihood of jumping at a certain site $\mathbf{r}_{new}(t)$, starting from current position $\mathbf{r}(t)$ (the current FOA). This can be evaluated as

$$p(s(\mathbf{r}_{new}(t))|s(\mathbf{r}_{(}t)),\mathbf{r}_{new}(t),\mathbf{r}(t))=\frac{\exp(-\beta_P(s(\mathbf{r}(t))-s(\mathbf{r}_{new}(t))))}{\sum_{\mathbf{r}'_{new}}\exp(-\beta_P(s(\mathbf{r}(t))-s(\mathbf{r}'_{new}(t))))}. \tag{6}$$

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.

The remaining factors in Eq. (5) summarize: (i) the motor action, where $p(\mathbf{r}_{new}(t)|\mathbf{r}(t),\theta,l)$ is the probability of the shift $\mathbf{r}(t)\rightarrow\mathbf{r}_{new}(t)$ given a pair (θ,l); (ii) the prior probabilities on action parameters (θ,l).

Following the discussion above, such priors are chosen as $p(\theta)=Unif(0,2\pi)$, the uniform distribution in the $[0,2\pi]$ interval and $p(l)=f(l;\alpha,\beta,\gamma,\delta)$.

By using Bayes' rule and Eq. (6), the choice of action parameters can thus be written as

$$\begin{aligned}&\arg\max_{\alpha,l}p(\theta,l|s(\mathbf{r}_{new}(t)),s(\mathbf{r}_{(}t)),\mathbf{r}_{new}(t),\mathbf{r}(t))\\&\quad=\arg\max_{\alpha,l}p(s(\mathbf{r}_{new}(t))|s(\mathbf{r}_{(}t)),\mathbf{r}_{new}(t),\mathbf{r}(t))\\&\quad p(\mathbf{r}_{new}(t)|\mathbf{r}(t),\theta,l)p(\theta)p(l).\end{aligned} \tag{7}$$

Eq. (7) can be evaluated through simple ancestral sampling on the generative model

$$\theta_k\sim Unif(0,2\pi),\quad k=1,\ldots,K, \tag{8}$$

$$l_k\sim f(l_k;\alpha,\beta,\gamma,\delta), \tag{9}$$

$$\mathbf{r}_{new,k}(t)\sim p(\mathbf{r}_{new}(t)|\mathbf{r}(t),\theta_k,l_k), \tag{10}$$

and by weighting the samples through the likelihood specified in Eq. (6).

Note that Eqs. (8)–(10) altogether provide a set of K motor actions that *a priori* could be undertaken by the forager. In particular Eq. (10), since representing the shift $\mathbf{r}(t) \rightarrow \mathbf{r}_{new}(t)$, can be implemented via Eq. (1) (in the molecular dynamics literature, this approach is known as Langevin Monte Carlo [42]).

In other terms, the motor step specified through Eq. (1) is used at this stage as an internal forward model [20,29], to simulate possible candidate flights among which the most likely actual flight is eventually determined by selecting the most suitable flight parameters by Eq. (7).

3.3. Intensive stage and behavioral switching

In the intensive stage the motion of the walker $\mathbf{r}(t) \rightarrow \mathbf{r}_{new}(t)$ accounts for both smooth pursuit and fixations.

Smooth pursuit allows smooth tracking of selected, and typically foveal, targets and represents an active response modulated by available cues (e.g., motion), attention, expectations [3]. Attention and pursuit share resources, and the selection of a pursuit target critically relies on attention. Indeed, after making a saccade to a moving object there is a high probability that pursuit will start [2], and moment-to-moment relocation of gaze is employed for the tracking. Only when abrupt changes of speed of directions occur, the walker might switch to execute saccades, switching to extensive relocation.

Fixations themselves are not simply the maintenance of the visual gaze on a single location but rather a slow oscillation of the eye [3]. 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° degrees of visual angle [43]).

For both mechanism, we let the intensive stage rely on a simple mechanism [44]: 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 [14].

Thus, the gaze moves directly to $\mathbf{r}_{new}(t)$, via, for instance, a winner takes all mechanism like the one proposed in [45]. In other words, let ρ be an arbitrary positive number; define $\mathbf{r}^*(t)$ as

$$\mathbf{r}^*(t) = \arg\max_{\mathbf{r}'(t)} \{s(\mathbf{r}'(t))\}_{\mathbf{r}'(t) \in \mathcal{N}_{\mathbf{r}(t)}}, \tag{11}$$

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

Eventually, the activities performed along the extensive and intensive stages, together with the mechanism for switching between the two can be formalized as follows.

Let $\Delta s = s(\mathbf{r}^*(t)) - s(\mathbf{r}(t))$ be the saliency gain and $\epsilon > 0$ an arbitrary threshold. The next position $\mathbf{r}_{new}(t)$ of gaze at time t, according to the composite foraging strategy defined by Eqs. (1) and (11), can be computed through the system of equations

$$\begin{cases} \mathbf{r}_{new}(t) = \xi \mathbf{r}^*(t) + (1-\xi)[\mathbf{r}(t) - \nabla V + \boldsymbol{\eta}] \\ \xi = H(\Delta s - \epsilon), \end{cases} \tag{12}$$

where H is the Heaviside function.

If $\Delta s > \epsilon$ then $\xi = 1$ and the foraging eye is in the intensive stage; if $\xi = 0$, extensive search is performed.

Finally, 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 [2,3]. In this perspective, the plausible center of a new fixation $\mathbf{r}_{new}(t)$, should be eventually accepted on the basis of some decision function $\mathcal{D}(\mathbf{r}_{new}(t), T)$, 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 will be described in the following section.

4. Simulations

4.1. Dataset

For simulations, we mainly used the CRCNS eye-1 dataset created by University of South California. The dataset is freely available http://crcns.org/data-sets/eye/eye-1 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). It comprises eye movement recordings from eight distinct subjects (three females and five males, ages 23–32, normal or corrected-to-normal vision) watching 50 different video clips (MPEG-1, 640 × 480 pixels, 30 fps, mean screen luminance 30 cd/m^2, room 4 cd/m^2, viewing distance 80 cm, field of view $28^\circ \times 21^\circ$, approximately 25 min of total playtime, 4–6 observers for each clip; 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 [46] for a description and https://crcns.org/files/data/eye-1/crcns-eye1-summary.pdf for more details.

The task given to the observers was: “Follow main actors and actions, try to understand overall what happens in each clip. We will ask you a question about main contents. Do not worry about details like specific text messages”.

Such task is methodologically compliant with the assumptions behind our study (see discussion in Section 2). On the one hand, the given task does not explicitly bias subjects toward low-level salient image locations; further, it avoids instructionless free viewing that often yields largely idiosyncratic patterns of eye movements. On the other hand, given all the factors behind the control of eye movements, it is still possible to evaluate the extent to which a model such as the one presented here, devoid of explicit notions of actors or actions, may yield a selection of scene locations that is comparable to the selection operated by human subjects.

4.2. Implementation details

Foveated input. Following [45], the FOA size is set as $|FOA| = 1/6 \min\{L,W\}$, L, W denoting the input frame height and width. Given a fixation point $\mathbf{r}_{new}(t)$ 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 isotropic 2-D Gaussian function centered at $\mathbf{r}_{new}(t)$ (the standard deviation is set to $|FOA|$), so to obtain the foveated frame $\widehat{\mathbf{I}}(t)$.

Saliency map. The foveated frame is used to compute the spatio-temporal saliency s through the self-resemblance method [47]. First a local image structure at each pixel is represented by a matrix of local descriptors (local regression kernels), which are robust in the presence of noise and image distortions. Then, matrix cosine similarity is employed to measure the resemblance of each pixel to its surroundings. For each pixel, the resulting saliency map represents the statistical likelihood of its feature matrix given the feature matrices of the surrounding pixels [47]. The saliency map is then normalized within the [0,100] range.

We initially experimented with different saliency computation methods [45,46,48]. However, self-resemblance provides comparable performance, but it can handle both static and space-time saliency computation, thus avoiding explicit motion estimation while being able to handle camera motion (see [47], for further details).

External potential. From s, landscape potential V) is computed via Eq. (3), with $\tau_V = 0.01$; then $\nabla V = [\partial V_x, \partial V_y]^\top$ is obtained using a finite difference method based on a central difference scheme.

Intensive stage. The direct vision range ρ, namely the radius of the circle $\mathcal{N}_{\mathbf{r}(t)}$, Eq. (11) is set equal to dimension of the FOA, $|FOA|$. Within such range $\mathbf{r}^*(t)$ is computed via Eq. (11) and the difference of saliency $\Delta s = s(\mathbf{r}^*(t)) - s(\mathbf{r}(t))$ evaluated and compared with threshold ϵ so to set the switching variable ξ in Eq. (12); ϵ has been experimentally determined as $\epsilon = 0.7 \max\{s(t)\}$.

Extensive stage. For what concerns the stochastic component, the optimal l,θ components to be chosen according to the MAP rule, Eq. (4), are obtained in practice via ancestral sampling, Eqs. (8)–(10), with $K=100$; $\beta_P = 1$ is used in Eq. (6).

The actual values of the motor parameters $\{\alpha,\beta,\gamma,\delta\}$ to be used in the sampling step of Eq. (9) 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.

Given the empirical distributions of gaze shifts (for more details see Section 4.3), 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. [49–51]. Based on these approximations, parameter estimation is facilitated using the estimator proposed in [50]. 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. (1).

Having fixed the parameters of the α-stable distribution, an α-stable random variable l_k can be sampled, Eq. (10), in several ways. The one applied here is the well known Chambers, Mallows, and Stuck procedure [52] (cf., Appendix A).

Flight decision. A straightforward computational implementation of the decision function $\mathcal{D}(\mathbf{r}_{new}(t),T)$ is given in the form of an acceptance process, implemented by a Metropolis algorithm, that again depends on the gain of saliency Δs and upon the "internal temperature" T, whose values determine the amount of randomness in the acceptance process: for low T values we will have a "quiet forager" behavior, whilst for higher values we will reach a higher randomness in scanpath generation. More precisely, the target site $\mathbf{r}_{new}(t)$ is accepted with probability (for simulations presented here $T=1000$)

$$p(a|\mathbf{r}(t)_{new}, \mathbf{r}(t)) = \min\{1, \exp(\Delta s/T)\}. \tag{13}$$

where a is a binary random variable denoting acceptance/rejection [42].

If no candidate FOA $\mathbf{r}(t)_{new}$ has been accepted, the current fixation point $\mathbf{r}(t)$ is maintained.

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, 4 GB RAM, under Mac OS X 10.5.8. As regards actual performance under such setting, the average elapsed time for the whole processing amounts to 2.179 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, 1.146 spf and saliency, 0.85 spf. The average elapsed time to compute the new point of gaze is 0.139 spf.

4.3. Results

We will discuss few examples that are representative of the results obtained on the CRCNS eye-1dataset.

The first example concerns the clip `beverly08` (see Fig. 1(b)). This sequence shows a daytime outdoor scene filmed at a park in Beverly Hills, recorded from a fixed camera. The scene background is initially static, then a jogger enters on the left and crosses the scene until he disappears to the right. Recorded eye traces show, as expected, variability in gaze patterns at the beginning and in the end of the clips, whilst, when the isolated jogger start to move a higher degree of interobserver coherence can be noticed [7].

Results of model simulation on this clip are summarized in Figs. 2–4. Each row of the figures presents the outputs of the main processing steps of the method previously

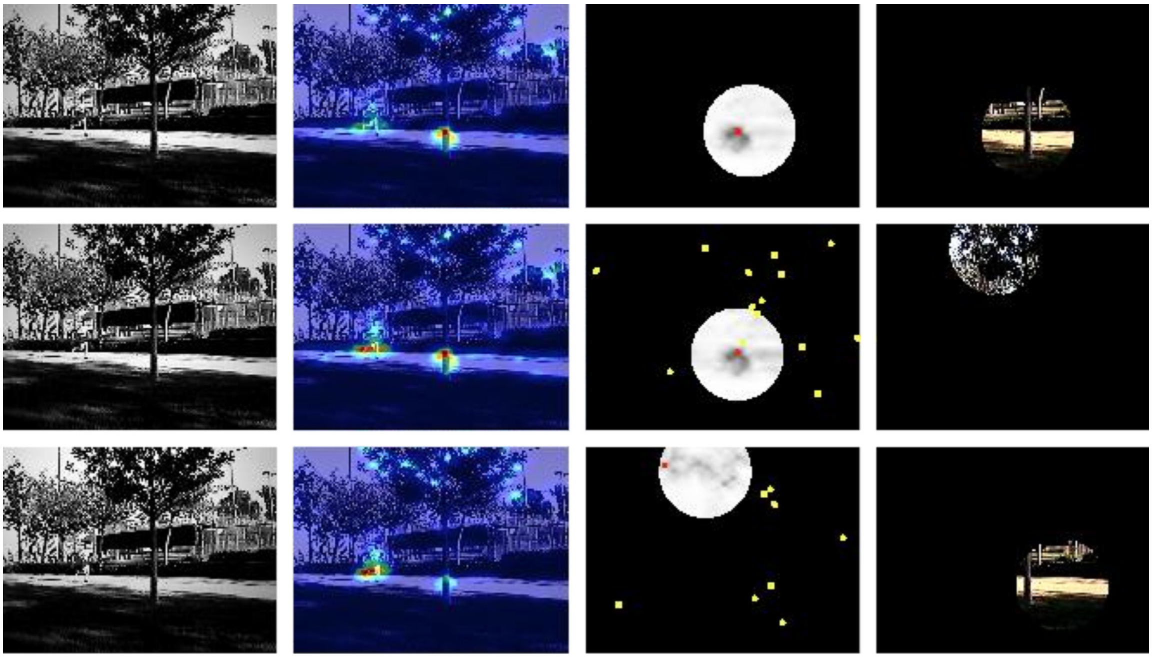

Fig. 2. Results obtained on three consecutive frames of the `beverly08` sequence. The first column (top to bottom) shows the sequence of foveated frames and the second column the corresponding saliency maps; the third column represents the direct vision area as a disk (red color representing maximum saliency) and yellow points indicate the simulated candidate FOAs; the fourth column shows the selected FOA.

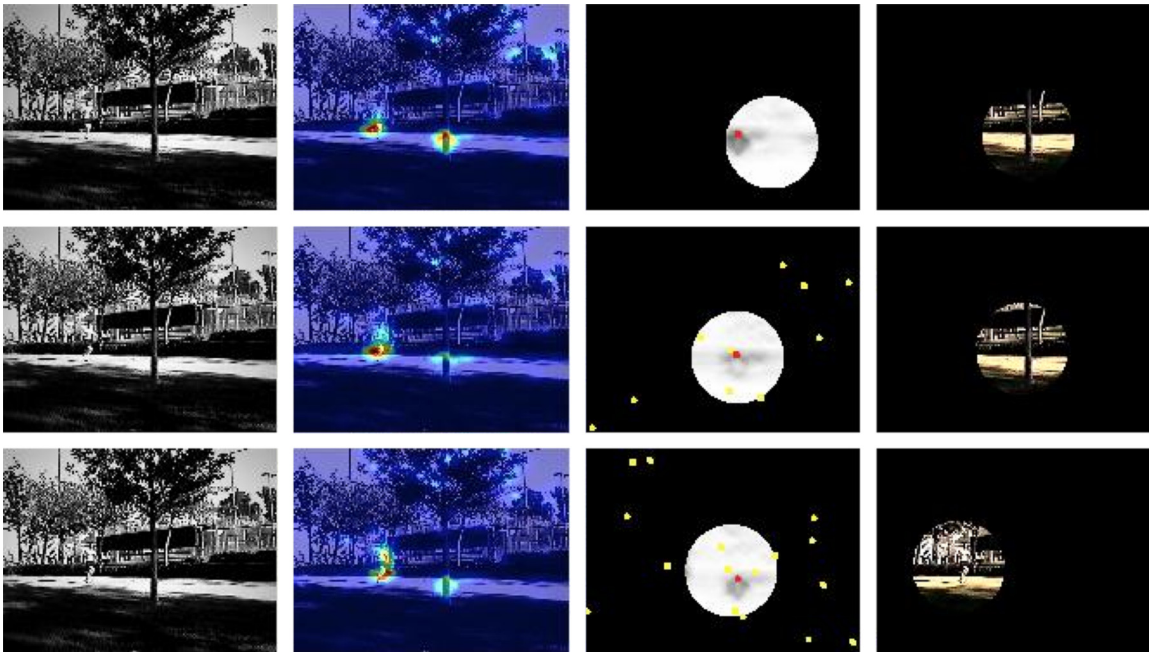

Fig. 3. Results obtained on the next three consecutive frames of the `beverly08` sequence. Organization of the figure is the same as in Fig. 2.

described: frame foveation, saliency extraction, proposal of candidate FOAs, gaze shifting. In particular, the third image within each row represents the direct vision area as a disk with a red spot standing for the point of maximum saliency (or minimum potential); when the random component is accounting for the sampling of candidate FOAs, these are represented as yellow points; clearly, the absence of yellow spots indicate that a candidate FOA has been successfully generated by the deterministic component of Eq. (12).

Consider first Fig. 2. After showing a fixation on the most salient object, the tree, global relocation of gaze is through Lévy flights. This behavior is prevalent in the initial

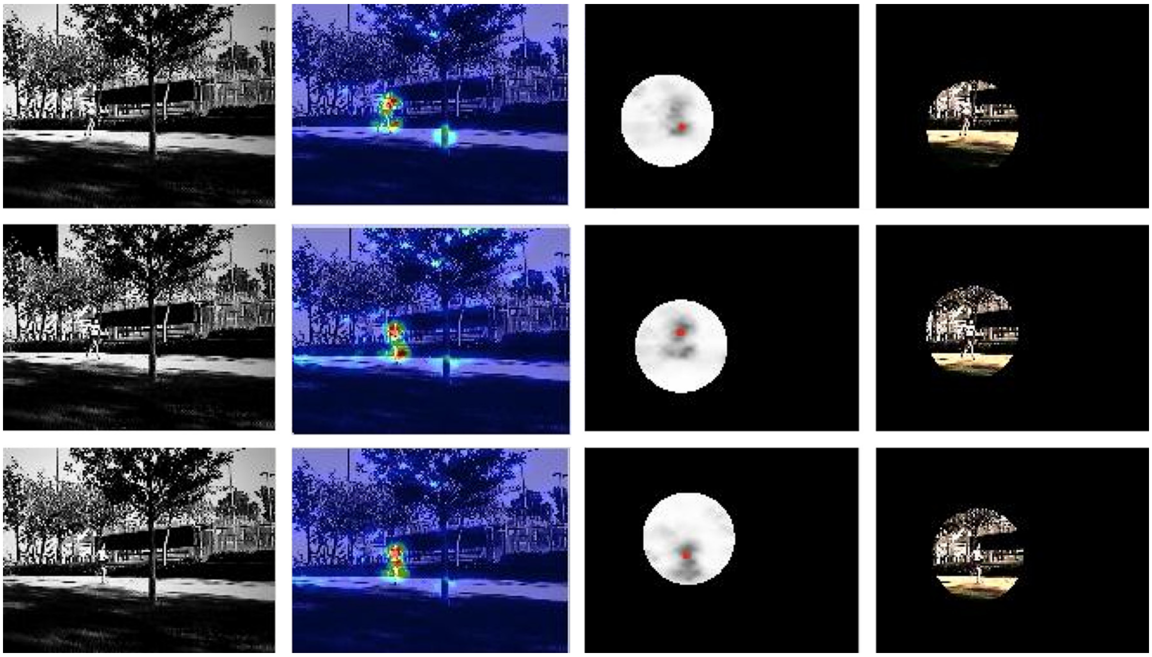

Fig. 4. Results obtained on the next three consecutive frames of the beverly08 sequence. Organization of the figure is the same as in Fig. 2.

part of the beverly08 clip, where due to the static background display, devoid of moving objects, gaze behavior is somehow similar to gaze behavior in picture viewing.

In subsequent frames (Fig. 3), the interplay between intensive and extensive stages gives as a result a fixation on the tree; in the meanwhile, the running person becomes more evident, and low-level saliency start to correlate to the moving objects. Then, through a medium saccade toward the left (bottom row in the same figure), attention is deployed to the runner.

This triggers a smooth pursuit phase (Fig. 4) mainly controlled through the intensive mechanism, which endures through time until the runner exits the scene.

To achieve some general insight on the behavior of the model, such results should be compared against those summarized in Fig. 5. The latter are obtained from the beverly01 clip. This sequence still shows a daytime outdoor scene filmed in Beverly Hills, but it has a higher degree of complexity than the beverly08 sequence: the distinction foreground/background is fuzzy, different kinds of motion are taking place either due to objects and people populating the scene, and to the mobile camera, yet zooming in the final part of the sequence.

One distinctive feature of the behavior of the system, which is readily apparent along the process of choosing where to look next, is that for the beverly08 sequence, the somehow regular structure of the observed dynamics reflects in an almost regular pattern of system behavior. The deterministic component of the system is prevalent in the occurrence of a unique object or action pattern (by fixation or smooth pursuit, e.g. gazing at jogging person), while the extensive stage mainly accounts for global scanning of the static background in the absence of moving objects/actions (initial and final parts of the clip).

As opposed to such behavior, in the beverly01 clip the complexity of the scene elicits a predominance of the extensive search mechanism, which results in a spatio-temporal scanpath characterized by a higher number of saccades.

For the purposes of this work, it is not an easy task to provide a measure of agreement between our models predictions and the scanpaths recorded from human observers. Here the issue is neither an evaluation of agreement with saliency measures nor can be reduced to requiring that humans fixated a given location at the same time [53] or within the same spatio-temporal "window" [54] as the model did. Indeed, similarly to eye-tracked human observers, simulated observers (see Fig. 10 later on) may agree on the average on where to look, at least in the spatial dimension [46]. In other terms, here we are not actually involved in determining an average attentive observer [55].

More subtly, the aim of this work is to ascertain whether the implemented model is capable of reproducing a gaze shift motor behavior, which exhibits statistics similar to those exhibited by human observers. 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 [34].

Thus, we studied the distributions of gaze magnitudes by analyzing eye-tracking results collected in the CRCNS database. Indeed, the study of flight magnitude 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 [38,56], human mobility [57], statistical physics (cf. [58] for a broad survey). To this end gaze shift samples from all

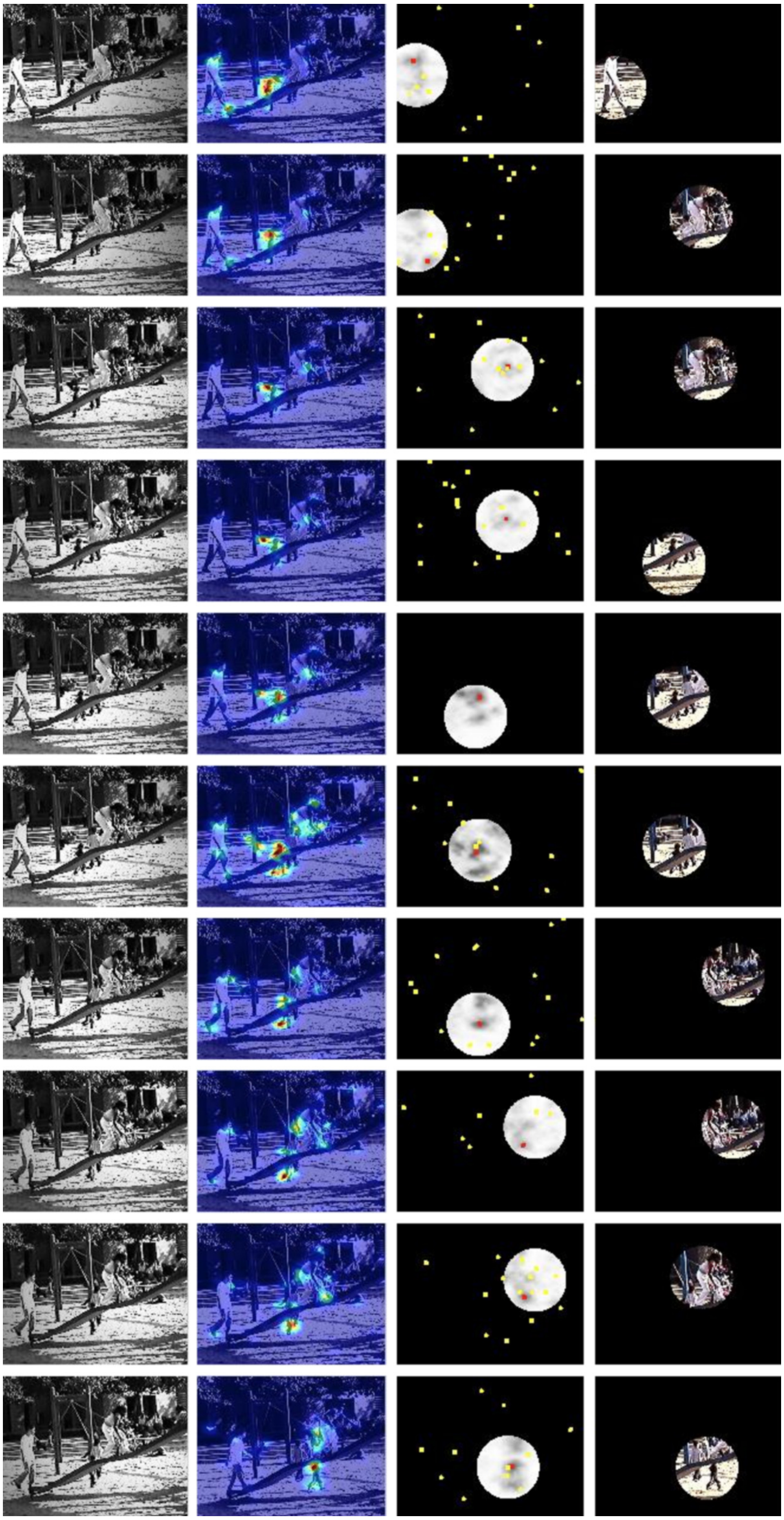

Fig. 5. An excerpt of results obtained on the `beverly01` clip from the CRCNS eye-1 dataset. The organization of the figure is the same of Fig. 2.

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 Lévy walks (e.g., [57]). In the CRCNS database, eye-tracker samples have been 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, since movies were displayed in the original experiment at a rate of 33.185 ms/frame, to be consistent, we subsample by 8 each smooth pursuit sub-tracks in order to work at a frame-rate basis, thus making it feasible to compare with the simulation.

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. The recorded FOA coordinates of the simulation have been used to compute the gaze magnitude empirical distributions (histograms) For a more precise description of the tail behavior of such distributions, 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).

Given the empirical distributions of eye-tracked and simulation gaze shifts obtained on the same video, the fit between the two is basically assessed via the two-sample Kolmogorov–Smirnov (K–S) test, which is very sensitive in detecting even a minuscule difference between two populations of data. We further check the results of the K–S test, using the Cramér-von Mises (C–M) test [59]. We also provide results from the standard Mann–Whitney U (MWU) test, to assess the null hypothesis that two samples have the same median (central tendency). All tests are performed at the level of significance $\alpha = 0.05$.

Fig. 6 shows results obtained on the `beverly08` clip in terms of both the gaze shift empirical magnitude distributions of eye-tracked subjects (Fig. 6(a)) and simulated

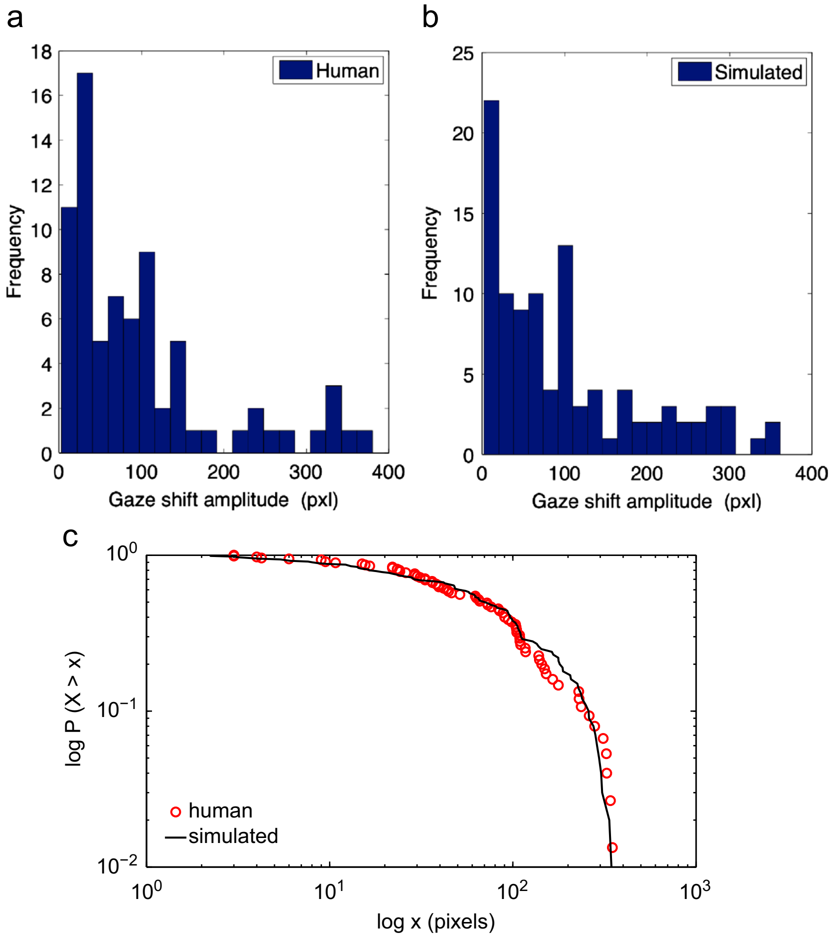

Fig. 6. Gaze shift magnitude distributions (eye-tracked on the left, simulated on the right) and CCDFs plotted on double log-scale for the `beverly08` clip. (a) Gaze shift magnitude distributions (human), (b) gaze shift magnitude distributions (model) and (c) complementary CDFs.

by the method (Fig. 6(b)) together with their corresponding upper tail behaviors (Fig. 6(c)). Algorithm generated scanpaths show similar gaze magnitude statistics described in terms of the complementary CDFs plotted on double log-scale. For this example (# samples = 175), the K–S test shows that there is no significant departure between simulated and eye-tracked data (K–S statistics = 0.083, *p*-value = 0.915672), and that is confirmed by the CvM test (C–M statistics = 0.065, *p*-value = 0.215964) and by MWU test (*p*-value = 0.951915).

These results hold also for the `beverly01` clip (see Fig. 7): again, from the K–S test the simulated distribution results no significantly different from the empirical one (# samples = 606, K–S statistics = 0.0815, *p*-value = 0.1121). The same holds for the CvM test (C–M statistics = 0.318, *p*-value = 0.879873) and the MWU test (*p*-value = 0.285078).

We illustrates results for other two clips, `monica03` and `tv-sports03`. The `monica03` clip concerns a urban traffic sequence, where attention will be attracted by traffic guards stationing at the crossing, pedestrian occasionally entering the scene, or vehicles (basically at constant velocity); it is comparable in complexity with `beverly01`. The `tv-sports03` clip is different in content from the others (displaying outdoor scenes) in that it shows a TV basketball game in which the dynamics is generated by pattern of sport actions, with occasional distractors such as advertisement, text, etc.

Results obtained for the `monica03` and `tv-sports03` clips are provided in Figs. 8 and 9 are similar to those obtained previously. For the `monica03` example (# samples = 2083) we have the K–S statistics = 0.0606, *p*-value = 0.074138; the result is confirmed by the CvM test (C–M statistics = 0.254, *p*-value = 0.816827). The assumption of equal central tendency is not rejected by the MWU test (*p*-value = 0.821786).

An analogous conclusion can be derived for `tv-sports03` (# samples = 1855, K–S test: K–S statistics = 0.0608, *p*-value = 0.111; CvM test: C–M statistics = 0.308, *p*-value = 0.871909; MWU test: *p*-value = 0.144653).

One interesting issue that could be raised here is how the model outputs, expressed in terms of saccade amplitude distributions, covaries with the different inputs (video saliency map streams). In other terms: are the observed amplitude distributions to be regarded as the output of a video-agnostic model, only fitting general oculomotor biases or do these also capture some characteristics of the underlying saliency distributions? Specific covariation can be appreciated at a glance in the case of the

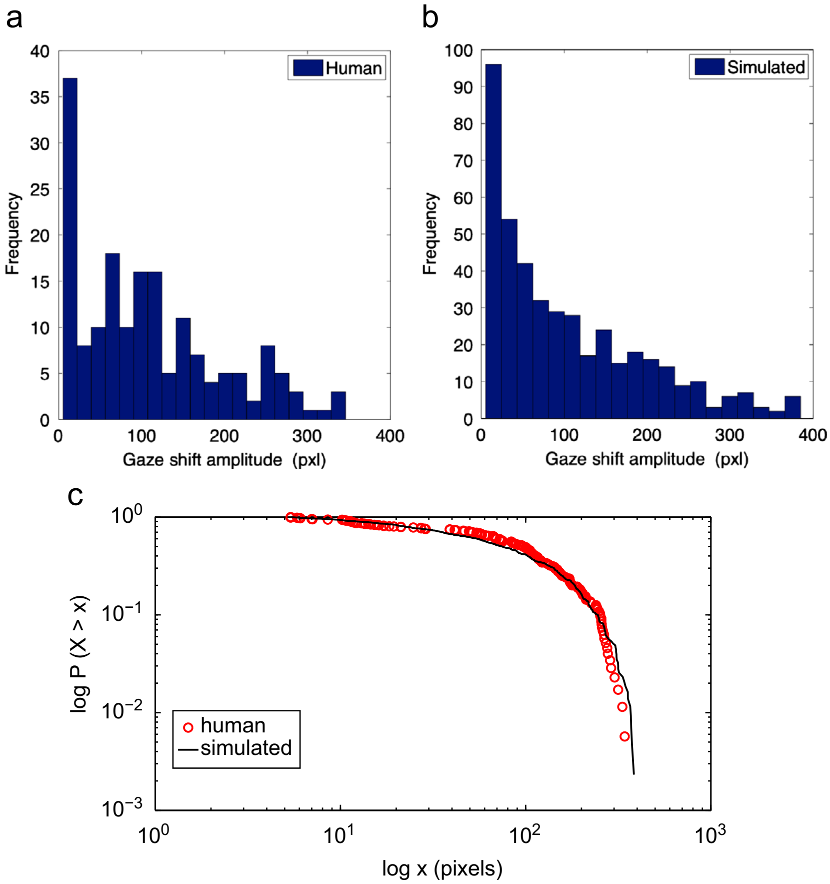

Fig. 7. Gaze shift magnitude distributions (eye-tracked on the left, simulated on the right) and CCDFs plotted on double log-scale for the `beverly01` clip. (a) Gaze shift magnitude distributions (human), (b) gaze shift magnitude distributions (model) and (c) complementary CDFs.

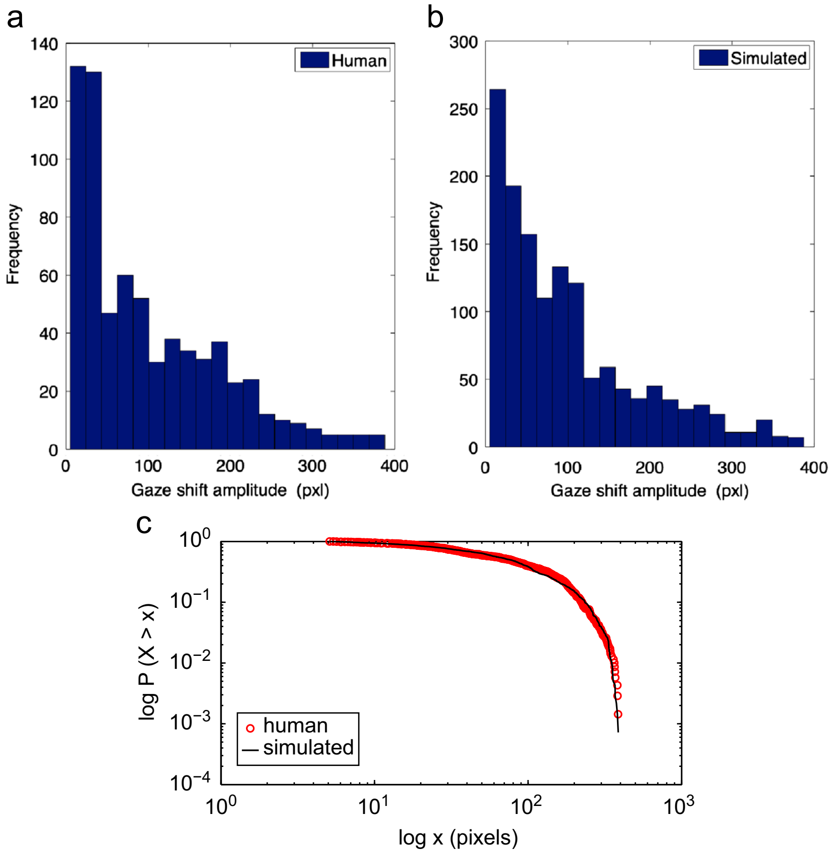

Fig. 8. Gaze shift magnitude distributions (eye-tracked on the top, simulated in the middle) and CCDFs plotted on double log-scale for the `monica03` clip. (a) Gaze shift magnitude distributions (human), (b) gaze shift magnitude distributions (model) and (c) complementary CDFs.

`beverly08` clip, whose distributions look quite different from those of other movies. However, in general, visually assessing the impact of different movies/saliency distributions is far from evident.

To this end, the K–S test has been extended to cross-compare the model's output for one video against human data distributions gathered from other videos; results are presented in Table 1 (clearly, data reported in the diagonal entries of the table are those previously discussed). As it can be seen, for off-diagonal entries, the only case where there is no statistically significant difference between amplitude distributions is when comparing results obtained from clips `beverly01` and `monica03`. Summing up, while capturing systematic tendencies in the form of long-tailed distributions [24], yet the model somehow accounts for peculiar characteristics of the input video, although most remarkable variations in distribution shape are likely to be achieved when changing the assigned task [10].

Finally, coming back to the original motivation of the model, mimicking the variability in gaze shifts, the example in Fig. 10 shows the individual gaze shifting behavior of three different "observers" simulated by running the method three times on an excerpt of the `beverly01` clip, but holding the same parameters that were used to generate previous figures.

5. Discussion and conclusions

In this study we have presented a stochastic model of eye guidance on complex videos, where the eye's behavior is that of an information random walker following a composite foraging strategy. The main goal of such model is to account for and mimic the variability of gaze shift patterns and their statistical properties as closely as possible.

Indeed, taking into account the randomness of the process may be relevant in computer vision and learning tasks [60,13], to provide principled visual attention models for video coding [61,53,62], image/video retrieval domains [63], and integrating the human attention analysis into video quality assessment (see [64] for a broad survey).

The variability issue has been surprisingly overlooked by most computational models of visual attention (cf., [4]), with few notable exceptions. In [65,66] simple eye-movements patterns are incorporated as a prior of a Dynamic Bayesian Network to guide the sequence of eye focusing positions on videos. The model presented in [67] embeds at least one parameter suitable to be tuned to obtain different saccade length distributions on static images. Closer to our study is the model by Keech and Resca [68] that mimics phenomenologically the eye movement trajectories

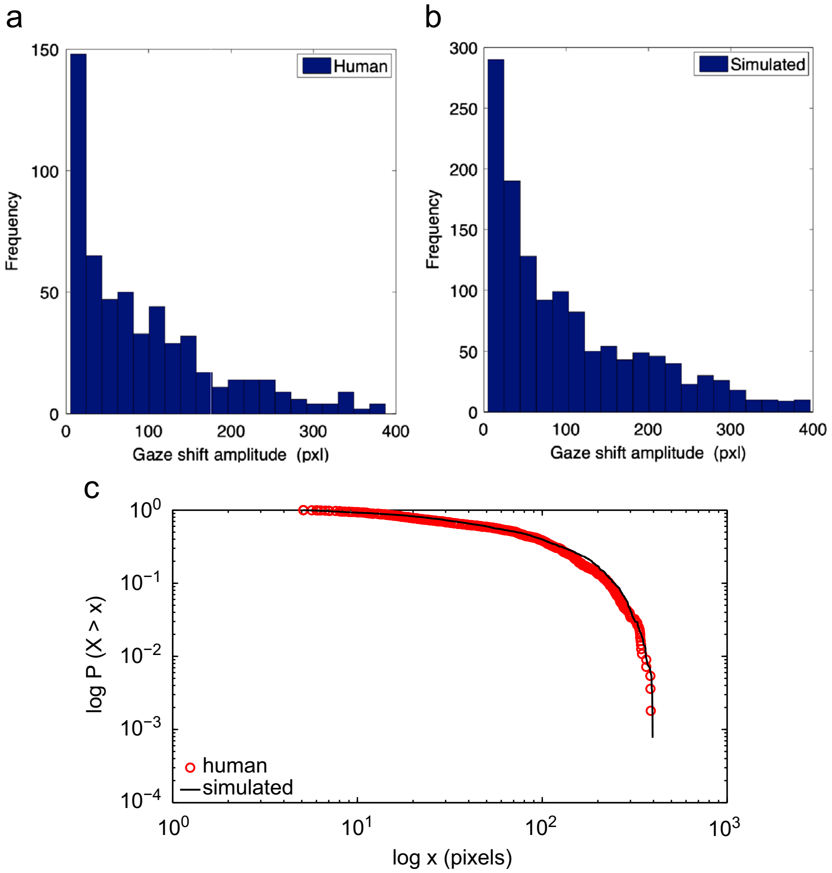

Fig. 9. Gaze shift magnitude distributions (eye-tracked on the left, simulated on the right) and CCDFs plotted on double log-scale for the `tv-sports03` clip. (a) Gaze shift magnitude distributions (human), (b) gaze shift magnitude distributions (model) and (c) complementary CDFs.

Table 1
K–S test results between simulation-generated (Model) and empirical (Human) distributions for the videos discussed in this paper. $H=0$ specifies that there is no significant departure between the two considered distributions (null hypothesis), otherwise $H=1$ holds. K–S statistics and p-values are reported in parenthesis.

Model	Human			
	`beverly01`	`beverly08`	`monica03`	`tv-sports03`
`beverly01`	$H=0$ ($KS=0.081$, $p=0.112$)	$H=1$ ($KS=0.175$, $p=0.032$)	$H=0$ ($KS=0.101$, $p=0.078$)	$H=1$ ($KS=0.173$, $p=0.026$)
`beverly08`	$H=1$ ($KS=0.171$, $p=0.046$)	$H=0$ ($KS=0.083$, $p=0.915$)	$H=1$ ($KS=0.189$, $p=0.003$)	$H=1$ ($KS=0.175$, $p=0.011$)
`monica03`	$H=0$ ($KS=0.101$, $p=0.009$)	$H=1$ ($KS=0.19$, $p=0.007$)	$H=0$ ($KS=0.06$, $p=0.074$)	$H=1$ ($KS=0.142$, $p=0.001$)
`tv-sports03`	$H=1$ ($KS=0.127$, $p=0.021$)	$H=1$ ($KS=0.183$, $p=0.01$)	$H=1$ ($KS=0.096$, $p=0.003$)	$H=0$ ($KS=0.06$, $p=0.111$)

observed in a conjunctive visual search task and where randomness is captured in the model through a Monte Carlo selection of a particular eye movement based on its probability. Probabilistic modeling of eye movement data during a conjunction search task is also discussed in [69]. Other exceptions are given, but in the very specific field of eye-movements in reading [34].

Indeed, many computational models basically resort to deterministic mechanisms to realize gaze shifts, and curiously this has been the main route for modelling the most random kind of gaze shifts, saccades. Thus they will basically generate the same scanpath, in presence of the same input saliency map. Further, overlooking motor strategies and tendencies that characterize gaze shift programming results in distributions of saccade amplitudes that do not match human eye behavior, failing to account properly for the characteristic positively skewed distribution of amplitudes. For one clear example, confronting human generated distributions with those produced by the well-known Itti, Koch and Niebur's scheme [45]—probably the most adopted in actual systems, see [9].

Certainly, there are models of visual attention and gaze shifting that have been conceived in a statistical framework (for instance, see the section on Bayesian Models in the Borji

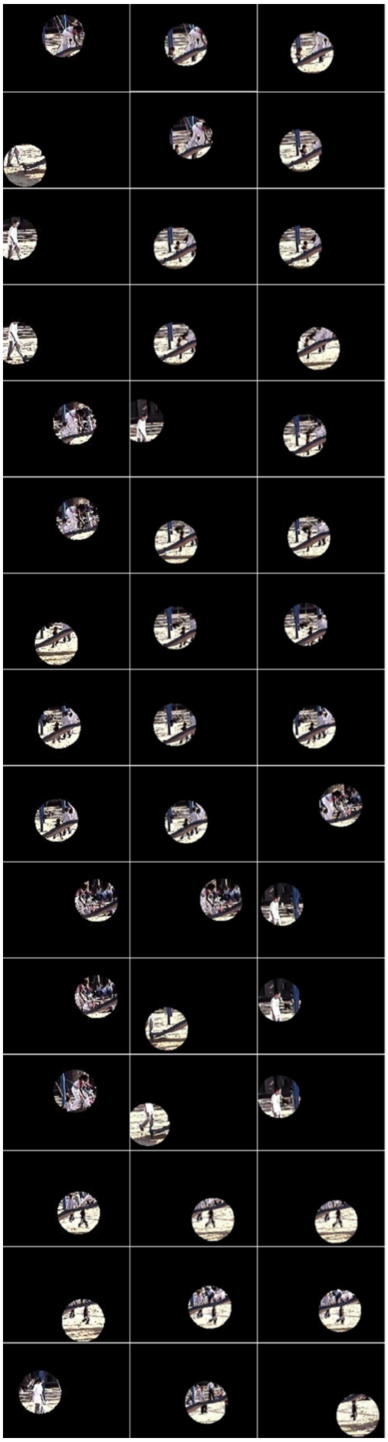

Fig. 10. An example of gaze shift dynamics by three different simulated observers. Each column shows results obtained on the same frames from the beverly01 clip illustrated in Fig. 5.

and Itti review [4], or, more specifically, [46,70–72]); nevertheless, even when probabilistic frameworks are used to infer where to look next, the final decision is often taken according to some "normalizing" rule (e.g., the MAP rule).

As a more general remark, up to this date, most computational models of visual attention are incomplete in the sense that the scope of their investigation is focused on the representational issues (low-level salience, top-down or context modulated salience, etc.), rather than considering the mechanisms for actually generating the sequence of gaze shifts.

As an effort in such direction, the model presented here extends previous ones [16,17,14] in two ways: by providing a composition or layering of strategies and by exploiting a sort of internal model akin to provide, through simulation, a more effective and efficient way of visually sampling the environment.

The composition of strategies allows to treat different types of eyes movement (smooth pursuit/saccade) within the same framework and makes this approach a step towards the unified modelling of different kinds of gaze shifts [3]; indeed, there is growing evidence that pursuit and saccades are interwoven to provide eye guidance, and are not completely independent systems as assumed for long time [2].

In addition, it stresses the importance of the role of the motor component, which is often neglected in the literature [73]. Further, this approach may be developed for a principled modelling of individual differences, a key issue in cognitive science [74], since providing cues for defining the informal notion of scanpath idiosyncrasy in terms of individual gaze shift distribution parameters [69].

The internal simulation step, beyond bearing a relation to motor programming theories [20–22], by providing an informed mechanism of choice has also connections with visibility or value models [2,3]. In this perspective, eye movements serve to provide new information about the surroundings, maximizing information gathering or reducing uncertainty about the visual stimulus [75].

Note that, even though the experiments presented here relied upon bottom-up cues, since for computational efficiency the saliency step was based on the method described in [47], the model is not concerned with such limitation. Nothing prevents from adopting a top-down derived saliency map (e.g., [70]). 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 at regions representing objects or proto-objects [76] that have relevance in determining organism behavioral responses.

One limitation of the model is that currently, the prior distribution of gaze relocation directions is chosen as uniform, whilst it has been reported that the distribution of saccade angles is skewed towards the horizontal and vertical axes [23,24]. This may be due to a cost minimization strategy (oblique movements are more complex) in which case this bias should be incorporated in the prior $p(\theta)$. By contrast, studies of natural scene statistics show that there is often most power in the horizontal directions, followed by vertical and finally oblique directions [77,78]. Thus, the skewness in direction distributions may reflect the posterior distribution of θ rather than the prior. Further, there is evidence for memory and persistence effects in directions (forward bias, [68]), that may be better modelled by considering a conditional prior in the form $p(\theta(t)|\theta(t-1))$.

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.

Appendix A. α-Stable distributions

The class of α-stable distributions [15] 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 $E[\exp(itx)]=\int_{\mathbb{R}}\exp(itx)\,dF(x)$, F being the cumulative distribution function (CDF), can be written as

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

the first expression holding if $\alpha\neq1$, 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 [52,49,50].

The technique for sampling $x\sim f(x;\alpha,\beta,\gamma,\delta)$ used in the present study is the well known Chambers, Mallows, and Stuck procedure [52]:

1. Generate a value Z from a standard stable $f(\alpha,\beta,0,1)$:

$$V\sim Unif\left(-\frac{\pi}{2},\frac{\pi}{2}\right); \tag{A.1}$$

$$W\sim\exp(1). \tag{A.2}$$

where $Unif(\cdot)$ denotes the uniform distribution.

2. If $\alpha\neq1$:

$$Z=S_{\alpha,\beta}\frac{\sin(\alpha(V+B_{\alpha,\beta}))}{\cos(V)^{1/\alpha}}\left(\frac{\cos(V-\alpha(V+B_{\alpha,\beta}))}{W}\right)^{(1-\alpha)/\alpha}, \tag{A.3}$$

where $S_{\alpha,\beta}=\arctan(\beta\tan(\pi\alpha/2))/\alpha$ and $B_{\alpha,\beta}=(1+\beta^2\tan^2(\pi\alpha/2))^{1/2\alpha}$.

3. If $\alpha=1$, then

$$Z=\frac{2}{\pi}\left[\left(\frac{\pi}{2}+\beta V\right)\tan(V)-\beta\log\left(\frac{W\cos(V)}{\frac{\pi}{2}+\beta V}\right)\right]. \tag{A.4}$$

Once a value Z from a standard stable $f(\alpha,\beta,0;1)$ has been simulated, in order to obtain a value x from a distribution with scale parameter γ and location parameter δ, the following transformation is required: $x=Z+\delta$ if $\alpha\neq1$; $x=\gamma Z+(2/\pi)\beta\gamma\log(\gamma)+\delta$, if $\alpha=1$.

References

[1] D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, W.H. Freeman, New York, 1982.

[2] A. Schütz, D. Braun, K. Gegenfurtner, Eye movements and perception: a selective review, Journal of Vision 11 (5).

[3] E. Kowler, Eye movements: the past 25 years, Vision Research 51 (13) (2011) 1457–1483. 50th Anniversary Special Issue of Vision Research—Volume 2.

[4] A. Borji, L. Itti, State-of-the-art in visual attention modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence, http://dx.doi.org/10.1109/TPAMI.2012.89, in press.

[5] S. Frintrop, E. Rome, H. Christensen, Computational visual attention systems and their cognitive foundations: a survey, ACM Transactions on Applied Perception 7 (1) (2010) 6.

[6] M. Begum, F. Karray, Visual attention for robotic cognition: a survey, IEEE Transactions on Autonomous Mental Development 3 (1) (2011) 92–105.

[7] M. Dorr, T. Martinetz, K. Gegenfurtner, E. Barth, Variability of eye movements when viewing dynamic natural scenes, Journal of Vision 10 (10).

[8] B. Tatler, R. Baddeley, I. Gilchrist, Visual correlates of fixation selection: effects of scale and time, Vision Research 45 (5) (2005) 643–659.

[9] B. Tatler, M. Hayhoe, M. Land, D. Ballard, Eye guidance in natural vision: reinterpreting salience, Journal of Vision 11 (5).

[10] R. Canosa, Real-world vision: selective perception and task, ACM Transactions on Applied Perception 6 (2) (2009) 11.

[11] C. Harris, On the optimal control of behaviour: a stochastic perspective, Journal of Neuroscience Methods 83 (1) (1998) 73–88.

[12] H. Martinez, M. Lungarella, R. Pfeifer, Stochastic Extension to the Attention-Selection System for the iCub, University of Zurich, Technical Report.

[13] Y. Nagai, From bottom-up visual attention to robot action learning, in: IEEE Eighth International Conference on Development and Learning, IEEE Press, 2009, pp. 1–6.

[14] G. Boccignone, M. Ferraro, Modelling eye-movement control via a constrained search approach, in: Proceedings of Third European Workshop on Visual Information Processing (EUVIP 2011), IEEE Press, 2011, pp. 235–240.

[15] B. Gnedenko, A. Kolmogórov, Limit Distributions for Sums of Independent Random Variables, Addison-Wesley Publishing Co., 1954.

[16] D. Brockmann, T. Geisel, The ecology of gaze shifts, Neurocomputing 32 (1) (2000) 643–650.

[17] G. Boccignone, M. Ferraro, Modelling gaze shift as a constrained random walk, Physica A: Statistical Mechanics and its Applications 331 (1–2) (2004) 207–218.

[18] A. Yarbus, Eye Movements and Vision, Plenum Press, New York, 1967.

[19] G. Buswell, How People Look at Pictures: A Study of the Psychology and Perception in Art, University Chicago Press, Chicago, 1935.

[20] D. Wolpert, M. Kawato, Multiple paired forward and inverse models for motor control, Neural Networks 11 (7–8) (1998) 1317–1329.

[21] K. Körding, D. Wolpert, Bayesian integration in sensorimotor learning, Nature 427 (6971) (2004) 244–247.

[22] J. Fuster, Upper processing stages of the perception—action cycle, Trends in Cognitive Sciences 8 (4) (2004) 143–145.

[23] B. Tatler, B. Vincent, Systematic tendencies in scene viewing, Journal of Eye Movement Research 2 (2) (2008) 1–18.

[24] B. Tatler, B. Vincent, The prominence of behavioural biases in eye guidance, Visual Cognition 17 (6–7) (2009) 1029–1054.

[25] B. Tatler, R. Baddeley, B. Vincent, The long and the short of it: spatial statistics at fixation vary with saccade amplitude and task, Vision Research 46 (12) (2006) 1857–1862.

[26] E. Over, I. Hooge, B. Vlaskamp, C. Erkelens, Coarse-to-fine eye movement strategy in visual search, Vision Research 47 (2007) 2272–2280.

[27] P. Unema, S. Pannasch, M. Joos, B. Velichkovsky, Time course of information processing during scene perception: the relationship between saccade amplitude and fixation duration, Visual Cognition 12 (3) (2007) 473–494.

[28] A. Faisal, L. Selen, D. Wolpert, Noise in the nervous system, Nature Reviews Neuroscience 9 (4) (2008) 292–303.

[29] P. Bays, D. Wolpert, Computational principles of sensorimotor control that minimize uncertainty and variability, The Journal of Physiology 578 (2) (2007) 387–396.

[30] R. Abrams, D. Meyer, S. Kornblum, Speed and accuracy of saccadic eye movements: characteristics of impulse variability in the oculomotor system, Journal of Experimental Psychology: Human Perception and Performance 15 (3) (1989) 529.

[31] C. Harris, D. Wolpert, The main sequence of saccades optimizes speed-accuracy trade-off, Biological Cybernetics 95 (1) (2006) 21–29.

[32] R. van Beers, The sources of variability in saccadic eye movements, The Journal of Neuroscience 27 (33) (2007) 8757–8770.

[33] C. Quaia, W. Joiner, E. FitzGibbon, L. Optican, M. Smith, Eye movement sequence generation in humans: motor or goal updating?, Journal of Vision 10 (14).

[34] G. Feng, Eye movements as time-series random variables: a stochastic model of eye movement control in reading, Cognitive Systems Research 7 (1) (2006) 70–95.

[35] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guérin-Dugué, Modelling spatio-temporal saliency to predict gaze direction for short videos, International Journal of Computer Vision 82 (3) (2009) 231–243.

[36] G. Viswanathan, E. Raposo, M. da Luz, Lévy flights and superdiffusion in the context of biological encounters and random rearches, Physics of Life Review 5 (3) (2008) 133–150.

[37] E. Codling, M. Plank, S. Benhamou, Random walk models in biology, Journal of the Royal Society Interface 5 (25) (2008) 813.

[38] M. Plank, A. James, Optimal foraging: Lévy pattern or process? Journal of the Royal Society Interface 5 (26) (2008) 1077.

[39] A. Reynolds, Optimal random Lévy-loop searching: new insights into the searching behaviours of central-place foragers, Europhysics Letters 82 (2008) 20001.

[40] A. Bahill, D. Adler, L. Stark, Most naturally occurring human saccades have magnitudes of 15 degrees or less, Investigative Ophthalmology & Visual Science 14 (6) (1975) 468–469.

[41] M. Land, N. Mennie, J. Rusted, The roles of vision and eye movements in the control of activities of daily living, Perception 28 (1999) 1311–1328.

[42] R. Neal, Probabilistic inference using Markov chain Monte Carlo methods, Department of Computer Science, University of Toronto, 1993. URL 〈http://www.cs.utoronto.ca/radford/〉.

[43] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, J. Van de Weijer, Eye Tracking: A Comprehensive Guide to Methods and Measures, Oxford University Press, Oxford, UK, 2011.

[44] M. Da Luz, S. Buldyrev, S. Havlin, E. Raposo, H. Stanley, G. Viswanathan, Improvements in the statistical approach to random Lévy flight searches, Physica A: Statistical Mechanics and its Applications 295 (1–2) (2001) 89–92.

[45] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 1254–1259.

[46] L. Itti, P. Baldi, Bayesian surprise attracts human attention, Vision Research 49 (10) (2009) 1295–1306.

[47] H. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, Journal of Vision 9 (12) (2009) 1–27.

[48] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Advances in Neural Information Processing Systems, vol. 19, MIT Press, Cambridge, MA, 2007, pp. 545–552.

[49] J. Nolan, Numerical calculation of stable densities and distribution functions, Communications in Statistics–Stochastic Models 13 (4) (1997) 759–774.

[50] I. Koutrouvelis, Regression-type estimation of the parameters of stable laws, Journal of the American Statistical Association (1980) 918–928.

[51] J. McCulloch, Simple consistent estimators of stable distribution parameters, Communications in Statistics: Simulation and Computation 15 (4) (1986) 1109–1136.

[52] J. Chambers, C. Mallows, B. Stuck, A method for simulating stable random variables, Journal of the American Statistical Association 71 (354) (1976) 340–344.

[53] L. Itti, Automatic foveation for video compression using a neurobiological model of visual attention, IEEE Transactions on Image Processing 13 (10) (2004) 1304–1318.

[54] G. Boccignone, A. Marcelli, P. Napoletano, G. Di Fiore, G. Iacovoni, S. Morsa, Bayesian integration of face and low-level cues for foveated video coding, IEEE Transactions on Circuits and Systems for Video Technology 18 (12) (2008) 1727–1740.

[55] O. Le Meur, A. Ninassi, P. Le Callet, D. Barba, Overt visual attention for free-viewing and quality assessment tasks: impact of the regions of interest on a video quality metric, Signal Processing: Image Communication 25 (7) (2010) 547–558.

[56] A. Reynolds, Lévy flight patterns are predicted to be an emergent property of a bumblebees foraging strategy, Behavioral Ecology and Sociobiology 64 (1) (2009) 19–23.

[57] I. Rhee, M. Shin, S. Hong, K. Lee, S. Kim, S. Chong, On the levy-walk nature of human mobility, IEEE/ACM Transactions on Networking 19 (3) (2011) 630–643.

[58] R. Metzler, 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 37 (2004) R161.

[59] T. Anderson, On the distribution of the two-sample Cramer–von Mises criterion, The Annals of Mathematical Statistics (1962) 1148–1159.

[60] Y. Nagai, Stability and sensitivity of bottom-up visual attention for dynamic scene analysis, in: Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE Press, 2009, pp. 5198–5203.

[61] J. Lee, F. De Simone, T. Ebrahimi, Efficient video coding based on audio–visual focus of attention, Journal of Visual Communication and Image Representation 22 (8) (2011) 704–711.

[62] Z. Wang, L. Lu, A.C. Bovik, Foveation scalable video coding with automatic fixation selection, IEEE Transactions on Image Processing 12 (2003) 1–12.

[63] C. Cotsaces, N. Nikolaidis, I. Pitas, Video shot detection and condensed representation. A review, IEEE Signal Processing Magazine 23 (2) (2006) 28–37.

[64] J. You, U. Reiter, M. Hannuksela, M. Gabbouj, A. Perkis, Perceptual-based quality assessment for audio–visual services: a survey, Signal Processing: Image Communication 25 (7) (2010) 482–501.

[65] A. Kimura, D. Pang, T. Takeuchi, J. Yamato, K. Kashino, Dynamic Markov random fields for stochastic modeling of visual attention, in: Proceedings of the 19th International Conference on Pattern Recognition, 2008, ICPR 2008, IEEE, 2008, pp. 1–5.

[66] K. Miyazato, A. Kimura, S. Takagi, J. Yamato, Real-time estimation of human visual attention with dynamic Bayesian network and mcmc-based particle filter, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2009, ICME 2009, IEEE, 2009, pp. 250–257.

[67] T. Ho Phuoc, A. Guérin-Dugué, N. Guyader, A computational saliency model integrating saccade programming, in: Proceedings of the International Conference on Bio-inspired Systems and Signal Processing, Porto, Portugal, 2009, pp. 57–64.

[68] T. Keech, L. Resca, Eye movements in active visual search: A computable phenomenological model, Attention, Perception, & Psychophysics 72 (2) (2010) 285–307.

[69] U. Rutishauser, C. Koch, Probabilistic modeling of eye movement data during conjunction search via feature-based attention, Journal of Vision 7 (6).

[70] S. Chikkerur, T. Serre, C. Tan, T. Poggio, What and where: a Bayesian inference theory of attention, Vision Research 50 (22) (2010) 2233–2247.

[71] M. Begum, F. Karray, G. Mann, R. Gosine, A probabilistic model of overt visual attention for cognitive robots, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 40 (5) (2010) 1305–1318.

[72] A. Torralba, Contextual priming for object detection, International Journal of Computer Vision 53 (2003) 153–167.

[73] O. Braddick, J. Atkinson, Development of human visual function, Vision Research 51 (13) (2011) 1588–1609.

[74] J. Vandekerckhove, F. Tuerlinckx, M. Lee, Hierarchical diffusion models for two-choice response times, Psychological Methods 16 (1) (2011) 44.

[75] J. Najemnik, W. Geisler, Optimal eye movement strategies in visual search, Nature 434 (7031) (2005) 387–391.

[76] M. Wischnewski, A. Belardinelli, W. Schneider, J. Steil, Where to look next? Combining static and dynamic proto-objects in a TVA-based model of visual attention, Cognitive Computation 2 (4) (2010) 326–343.

[77] A. Torralba, A. Oliva, Statistics of natural image categories, Network: Computation in Neural Systems 14 (3) (2003) 391–412.

[78] A. Torralba, A. Oliva, M. Castelhano, J. Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search, Psychological Review 113 (4) (2006) 766.