# Attentive monitoring of multiple video streams driven by a Bayesian foraging strategy

Paolo Napoletano, *Member, IEEE,* Giuseppe Boccignone and Francesco Tisato

***Abstract*—In this paper we shall consider the problem of deploying attention to subsets of the video streams for collating the most relevant data and information of interest related to a given task. We formalize this monitoring problem as a foraging problem. We propose a probabilistic framework to model observer's attentive behavior as the behavior of a forager. The forager, moment to moment, focuses its attention on the most informative stream/camera, detects interesting objects or activities, or switches to a more profitable stream.**

**The approach proposed here is suitable to be exploited for multi-stream video summarisation. Meanwhile, it can serve as a preliminary step for more sophisticated video surveillance, e.g. activity and behavior analysis. Experimental results achieved on the UCR Videoweb Activities Dataset, a publicly available dataset, are presented to illustrate the utility of the proposed technique.**

***Index Terms*—Multi-camera video surveillance; Multi-stream summarisation; Cognitive Dynamic Surveillance; Attentive vision; Activity detection; Foraging theory; Intelligent sensors**

## I. Introduction

THE volume of data collected by current networks of cameras for video surveillance clearly overburdens the monitoring ability of human viewers to stay focused on a task. Further, much of the data that can be collected from multiple video streams is uneventful. Thus, the need for the discovery and the selection of activities occurring within and across videos for collating information most relevant to the given task has fostered the field of multi-stream summarise.

At the heart of multi-stream summarisation there is a "choose and leave" problem that moment to moment an ideal or optimal observer (say, a software agent) must solve: choose the most informative stream; detect, if any, interesting activities occurring within the current stream; leave the handled stream for the next "best" stream.

In this paper, we provide a different perspective to such "choose and leave" problem based on a principled framework that unifies overt visual attention behavior and optimal foraging. The framework we propose is just one, but a novel, way of formulating the multi-stream summarisation problem and solution (see Section II, for a discussion).

In a nutshell, we consider the foraging landscape of multiple streams, each video stream being a *foraging patch*, and the ideal observer playing the role of the visual forager (cfr. Table I). According to Optimal Foraging Theory (OFT), a forager that feeds on patchily distributed preys or resources, spends its time traveling between patches or searching and handling food within patches [1]. While searching, it gradually depletes the food, hence, the benefit of staying in the patch is likely to gradually diminish with time. Moment to moment, striving to maximize its foraging efficiency and energy intake, the forager should make decisions: Which is the best patch to search? Which prey, if any, should be chased within the patch? When to leave the current patch for a richer one?

Here visual foraging corresponds to the time-varying overt deployment of visual attention achieved through oculomotor actions, namely, gaze shifts. Tantamount to the forager, the observer is pressed to maximize his information intake over time under a given task, by moment-to-moment sampling the most informative subsets of video streams. All together,

TABLE I
Relationship between Attentive vision and Foraging

| Multi-stream attentive processing | Patchy landscape foraging |
|---|---|
| Observer | Forager |
| Observer's gaze shift | Forager's relocation |
| Video stream | Patch |
| Proto-object | Candidate prey |
| Detected object | Prey |
| Stream selection | Patch choice |
| Deploying attention to object | Prey choice and handling |
| Disengaging from object | Prey leave |
| Stream leave | Patch leave or giving-up |

choosing the "best" stream, deploying attention to within-stream activities, leaving the attended stream, represent the unfolding of a dynamic decision making process. Such monitoring decisions have to be made by relying upon automatic interpretation of scenes for detecting actions and activities. To be consistent with the terminology proposed in the literature [2], an *action* refers to a sequence of movements executed by a single *object* ( e.g., "human walking" or "vehicle turning right"). An *activity* contains a number of sequential actions, most likely involving multiple objects that interact or co-exist in a shared common space monitored by single or multiple cameras (e.g., "passengers walking on a train platform and sitting down on a bench"). The ultimate goal of activity modelling is to understand *behavior*, i.e. the meaning of activity in the shape of a semantic description. Clearly, action/activity/behavior analysis entails the capability of spotting objects that are of interest for the given surveillance task.

Thus, in the work presented here the visual objects of interest occurring in video streams are the *preys* to be chased and handled by the visual forager. Decisions at the finer level of a single stream concern which object is to be chosen

P. Napoletano and F. Tisato are with the Dipartimento di Informatica, Sistemistica e Comunicazione, Universitá degli Studi Milano–Bicocca , Viale Sarca 336, 20126 Milano, Italy.

G. Boccignone is with the Dipartimento di Informatica, Universitá degli Studi Milano Statale, Via Comelico 39/41 Milano, 20135 Italy.

and analyzed (*prey choice and handling*, depending on task), and when to disengage from the spotted object for deploying attention to the next (*prey leave*).

The reformulation of visual attention in terms of foraging theory is not simply an informing metaphor. What was once foraging for tangible resources in a physical space became, over evolutionary time, foraging in cognitive space for information related to those resources [3], and such adaptations play a fundamental role in goal-directed deployment of visual attention [4]. Under these rationales, we present a model of Bayesian observer's attentive foraging supported by the perception/action cycle presented in Fig. 1. Building on the perception/action cycle, visual attention provides an efficient allocation and management of resources.

The cycle embodies two main functional blocks: the perceptual component and the executive control component. The perceptual component is in charge of "What" to look for, and the executive component accounts for the overt attention shifts, by deciding "Where and How" to look at, i.e., the actual gaze position, and thus the observer's Focus of Attention (FoA). The observer's perceptual system operates on information represented at different levels of abstraction (from raw data to task dependent information); at any time, the currently sensed visual stimuli depend on the oculomotor action or gaze shift performed either within the stream (*within-patch*) or across streams (*between-patch*). Based on perceptual inferences at the different levels, the main feedback information passed on to the executive component, or controller, is an index of stream quality formalized in terms of the configurational complexity of potential objects sensed within the stream.

A stream is selected by relying upon its pre-attentively sensed quality. Once within stream, the observer attentively detects and handles objects that are informative for the given task. Meanwhile, by intra-stream foraging, the observer gains information on the actual stream quality in terms of experienced detection rewards. Object handling within the attended stream occurs until a decision is made to leave for a more profitable stream. Such decision relies upon a Bayesian strategy, which is the core of this paper. The strategy extends to stochastic landscapes observed under incomplete information, the deterministic global policy derived from classic Charnov's Marginal Value Theorem (MVT, [5]), while integrating within-stream observer's experience.

By relying on such perception/action cycle, we assume that the deployment of gaze to one video frame precisely reflects the importance of that frame. Namely, given a number of video streams as the input, at any point in time, we designate the current gazed frame as the relevant video frame to be included in the final output summarisation. The output succinctly captures the most important data (objects engaged in actions) for the surveillance analysis task.

The idea of a layered framework for the control of gaze deployment, implementing a general perception/action loop is an important one in the visual attention literature (cfr., Schütz *et al.* [6] for a discussion) and it has been fostered by Fuster [7], [8]. Such idea together with the assumption that attention is algorithmic in nature and needs not to occupy a distinct physical place in the brain is germane to our theme.

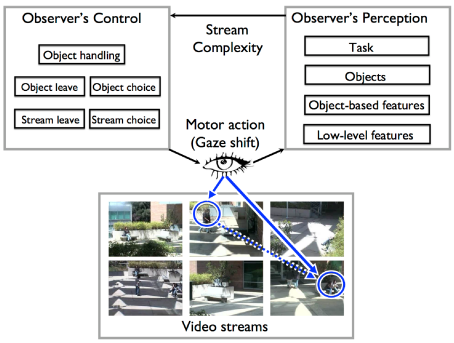


Fig. 1. Monitoring of multiple video streams as attentive foraging. The ideal observer (forager) is involved in a perception/action loop that supports foraging activity. Multiple streams are the raw sensory input. The observer pre-attentively selects the most informative stream (patch) and sets his Focus of Attention via a gaze shift action; within the stream, interesting objects (preys) are attentively detected and handled through local gaze shifts. Moment to moment, a Bayesian optimal strategy is exploited to make a decision whether to stay or leave the scrutinized stream by shifting the gaze to a more profitable one. The strategy relies upon the perceptual feedback of the overall "quality" (complexity) of streams.

Fuster's paradigm has been more recently formalized by Haykin [9] under the name of *Cognitive Dynamic Systems*. In this perspective, our cognitive foraging approach to monitoring can be considered closely related to the Cognitive Dynamic Surveillance System (CDSS) approach, a remarkable and emerging domain proposed by Regazzoni and colleagues [10], [11], [12]. In CDSS, attentive mechanisms [13] are likely to add relevant value in the effort of designing the next generation of surveillance systems.

In the rest of this paper, Section II describes the related literature and contributions of this work. Section III provides a formal overview of the model. Section IV details the pre-attentive stage. Stream selection is discussed in Section V. Section VI describes within-stream visual attention deployment, while Section VII discusses the Bayesian strategy for leaving the stream. Experimental work is presented in Section VIII. Finally, Section IX concludes this paper.

## II. Related work and our contributions

Main efforts in the summarisation literature have been spent on the single-camera case, while the multi-camera setting has not received as much attention (see [14], [15], [16] for review). Specifically, the work by Leo and Manjunath[15] shares our concern of providing a unified framework to generate summaries. Different from us, they rely on document analysis-inspired activity motif discovery. Time series of activations are computed from dense optical flow in different regions of the video and high-level activities are identified using the topic model analysis. The step from activities detected in individual video streams to a complete network summary relies on identifying and reducing inter-and intra-activity redundancy. This approach, cognate with those based on sparse coding dictionaries for finding the most representative frames (e.g.,[17]), requires off-line learning of activities from documents, each document being the time series of activations of a single stream. While offering some advantage for inferring high level activities[14], these methods avoid confronting with complex

vision problems and distributed optimal control strategies brought on by the multi-stream setting [18], [19], [20]. On the other hand, difficulties arise when dealing with large video corpora and with dynamic video streams, e.g. on-line summarisation in visual sensor networks [16], a case which is more related to our scenario.

In this view, beyond multi-camera summarisation, it is of interest work concerning multi-camera surveillance, where manual coordination becomes unmanageable when the number of cameras is large. To some extent, the "choice and leave" problem previously introduced bears relationships with two challenging issues: camera assignment (which camera is being used to extract essential information) and camera handoff (the process of finding the next best camera). Indeed, the complexity of these problems on large networks is such that Qureshi and Terzopoulos [21] have proposed the use of virtual environments to demonstrate camera selection and handover strategies. Two remarkable papers address the issue of designing a general framework inspired by non-conventional theoretical analyses, in a vein similar to the work presented here. Li and Bhanu [22] have presented an approach based on game-theory. Camera selection is based on a utility function that is computed by a bargaining among cameras capturing the tracking object. Esterle *et al.* [23] adopted a fully decentralised socio-economic approach for online handover in smart camera networks. Autonomous cameras exchange responsibility for tracking objects in a market mechanism in order to maximize their own utility. When a handover is required, an auction is initiated and cameras that have received the auction initiation try to detect the object within their the field of view.

At this point it is worth noting that, in the effort towards a general framework for stream selection and handling, all works above, differently from the approach we present here, are quite agnostic about the image analysis techniques to adopt. They mostly rely on basic tools (e.g., dense optical flow [15], Camshift tracking manually initialized [22], simple frame-to-frame SIFT computation [23]). However, from a general standpoint, moving object detection and recognition, tracking, behavioral analysis are stages that deeply involve the realms of image processing and machine vision. In these research areas, one major concern that has been an omnipresent topic during the last years is how to restrict the large amount of visual data to a manageable rate [19], [18].

Yet to tackle information overload, biological vision systems have evolved a remarkable capability: visual attention, which gates relevant information to subsequent complex processes (e.g., object recognition). A series of studies published under the headings of Animate [24], or Active Vision [25] has investigated how the concepts of human selective attention can be exploited for computational systems dealing with a large amount of image data (for an extensive review, see [26]). Indeed, determining the most interesting regions of an image in a "natural", human-like way is a promising approach to improve computational vision systems.

Surprisingly enough, the issue of attention has been hitherto overlooked by most approaches in video surveillance, monitoring and summarisation [14], [16], apart from those in the emerging domain of smart camera networks embedding pan-tilt-zoom (PTZ) cameras. PTZ cameras can actively change intrinsic and extrinsic parameters to adapt their field of view (FOV) to specific tasks [19], [20]. In such domain, active vision is a pillar [19], [18], since FOV adaptation can be exploited to focus the "video-network attention" on areas of interest. In PTZ networks, each of the cameras is assumed to have its own embedded target detection module, a distributed tracker that provides an estimate of the state of each target in the scene, and a distributed camera control mechanism [20]. Control issues have been central to this field: the large amount of camera nodes in these networks and the tight resource limitations requires balancing among conflicting goals [27], [21]. In this respect, the exploitation vs. exploration dilemma is cogent here much like in our work. For example, Sommerlade and Reid [28] present a probabilistic approach to maximize the expected mutual information gain as a measure for the utility of each parameter setting and task. The approach allows balancing conflicting objectives such as target detection and obtaining high resolution images of each target. Active distributed optimal control has been given a Bayesian formulation in a game theoretic setting. The Bayesian formulation enables automatic trading-off of objective maximization versus the risk of losing track of any target; the game-theoretic design allows the global problem to be decoupled into local problems at each PTZ camera [29], [30].

In most cases visual routines and control are treated as related but technically distinct problems [20]. Clearly, these involve a number of fundamental challenges to the existing technology in computer vision and the quest for efficient and scalable distributed vision algorithms [18]. The primary goal of these systems has been tracking distinct targets, where adopted schemes are extensions of the classic Kalman Filter to the distributed estimation framework [20]. However, it is important to note that tracking is but one aspect of multi-stream analysis and of visual attentive behavior ([2], but see Section III-B for a discussion). To sum up, while the development of PTZ networks has cast interest for active vision techniques that are at the heart of the attentive vision paradigm [24], [25], yet even in this field we are far from a full exploitation of tools made available by such paradigm.

There are some exceptions to this general state of affairs. The use of visual attention has been proposed by Kankanhalli *et al.*[31]. They embrace the broad perspective of multimedia data streams, but the stream selection process is yet handled within the classic framework of optimization theory and relying on an attention measure (saturation, [31]). Interestingly, they resort to the MVT result, but only for experimental evaluation purposes. In our work the Bayesian extension of the MVT is at the core of the process. The interesting work by Chiappino *et al.* [13] proposes a bio-inspired algorithm for attention focusing on densely populated areas and for detecting anomalies in crowd. Their technique relies on an entropy measure and in some respect bears some resemblance to the pre-attentive monitoring stage of our model. Martinel *et al.* [32] identify the salient regions of a given person, for person re-identification across non-overlapping camera views. Recent work on video summarisation has borrowed salience representations from the visual attention realm. Ejaz *et al.* [33]

choose key frames as salient frames on the basis of low-level salience. High-level salience based on most important objects and people is exploited in [34] for summarisation, so that the storyboard frames reflect the key object-driven events. Albeit not explicitly dealing with salience, since building upon sparse coding summarisation , Zhao and Xing [35] differentiate from [17] and generate video summaries by combining segments that *cannot* be reconstructed using the learned dictionary. Indeed, this approach, which incorporates in summaries unseen and interesting contents, is equivalent to denote salient those events that are unpredictable on prior knowledge (salient as "surprising", [26]). Either [34] and [35] only consider single-stream summarisation. The use of high-level saliency to handle the multi-stream case has been addressed in [36], hinging on [37]; this method can be considered as a baseline deterministic solution to the problem addressed here (cfr., for further analysis, Section VIII).

Our method is fundamentally different from all of the above approaches. We work within the attentive framework but the main novelty is that by focusing on the gaze as the principal paradigm for active perception, we reformulate the deployment of gaze to a video stream or to objects within the stream as a stochastic foraging problem. This way we unify intra- and inter-stream analyses. More precisely, the main technical contributions of this paper lie in the following.

First, based on OFT, a stochastic extension of the MVT is proposed, which defines an optimal strategy for a Bayesian visual forager. The strategy combines in a principled way global information from the landscape of streams with local information gained in attentive within-stream analysis. The complexity measure that is used is apt to be exploited for within-patch analysis (e.g, from group of people to single person behavior), much like some foragers do by exploiting a hierarchy of patch aggregation levels [38].

Second, the visual attention problem is formulated as a foraging problem by extending previous work on Lévy flights as a prior for sampling gaze shift amplitudes [39], which mainly relied on bottom-up salience. At the same time, task dependence is introduced, which is not achieved through ad hoc procedures. It is naturally integrated within attentional mechanisms in terms of rewards experienced in the attentive stage when the stream is explored. This issue is seldom taken into account in computational models of visual attention (see [26], [6] but in particular Tatler *et al* [40]). A preliminary study on this challenging problem has been presented in [41], but limited to the task of searching for text in static images.

## III. Model overview

In this Section we present an overview of the model to frame detailed discussion of its key aspects covered in Sections IV (pre-attentive analysis), V (stream choice), VI (within-stream attentive analysis) and VII (Bayesian strategy for stream leave).

Recall from Section I that the input to our system is a visual *landscape* of $K$ video streams, each stream being a sequence of time parametrized frames $\{\mathbf{I}^{(k)}(1), \mathbf{I}^{(k)}(2), \cdots, \mathbf{I}^{(k)}(t), \cdots\}$, where $t$ is the time parameter and $k \in [1, \cdots, K]$. Denote $\mathcal{D}$ the spatial support of $\mathbf{I}^{(k)}$, and $\mathbf{r}^{(k)} \in \mathcal{D}$ the coordinates of a point in such domain. By relying on the perception/action cycle outlined in Fig. 1, at any point $t$ in time, we designate the current gazed frame $\mathbf{I}^{(k)}(t)$ of stream $k$ as the relevant video frame to be selected and included in the final output summarisation

To such end, each video stream is the equivalent of a foraging patch (cfr. Table I) and objects of interest (preys) occur within the stream. In OFT terms, it is assumed that: the landscape is stochastic; the forager has sensing capabilities and it can gain information on patch quality and available preys as it forages. Thus, the model is conceived in a probabilistic framework. Use the following random variables (RVs):

- $\mathbf{T}$: a RV with $|\mathbf{T}|$ values corresponding to the *task* pursued by the observer.
- $\mathbf{O}$: a multinomial RV with $|\mathbf{O}|$ values corresponding to *objects* known by the observer

As a case study, we deal with actions and activities involving people. Thus, the given task $\mathbf{T}$ corresponds to "pay attention to people within the scene". To this purpose, the classes of objects of interest for the observer are represented by faces and human bodies, i.e., $\mathbf{O} = \{face, body\}$.

The observer engages in a perception/action cycle to accomplish the given task (Fig.1). Actions are represented by the moment-to-moment relocations of gaze, say $\mathbf{r}_F(t-1) \mapsto \mathbf{r}_F(t)$, where $\mathbf{r}_F(t-1)$ and $\mathbf{r}_F(t)$ are the old and new gaze positions, respectively. We deal with two kinds of relocations: i) from current video stream $k$ to the next selected $k'$ (between-patch shift), i.e. $\mathbf{r}_F^{(k)}(t-1) \mapsto \mathbf{r}_F^{(k')}(t)$; ii) from one position to another within the selected stream (within-patch gaze shifts), $\mathbf{r}_F^{(k)}(t-1) \mapsto \mathbf{r}_F^{(k)}(t)$. Since we assume unitary time for between-stream shifts, in the following we will drop the $k$ index and simply use $\mathbf{r}_F$ to denote the center of the FoA within the frame without ambiguity. Relocations occur because of decisions taken by the observer upon his own perceptual inferences. In turn, moment to moment, perceptual inferences are conditioned on the observer's current FoA set by the gaze shift action.

### A. Perceptual component

Perceptual inference stands on the visual features that can be extracted from raw data streams, a feature being a function $f : \mathbf{I}(t) \rightarrow F_f(t)$. In keeping with the visual attention literature [6], we distinguish between two kinds of features:

- *bottom-up* or feed-forward features, say $\mathbf{F}_{|\mathbf{I}}$ - such as edge, texture, color, motion features - corresponding to those that biological visual systems learn along evolution or in early development stages for identifying sources of stimulus information available in the environment (*phyletic* features, [8]);
- *top-down* or object-based features, i.e. $\mathbf{F}_{|\mathbf{O}}$.

There is a large variety of bottom-up features that could be used (see [26]). Following [42], we first compute, at each point $\mathbf{r}$ in the spatial support of the frame $\mathbf{I}(t)$ from the given stream, spatio-temporal first derivatives (w.r.t temporally adjacent frames $\mathbf{I}(t-1)$ and $\mathbf{I}(t+1)$). These are exploited to estimate, within a window, local covariance matrices $\mathbf{C}_\mathbf{r} \in \mathbb{R}^{3\times3}$, which in turn are used to compute space-time local

steering kernels $K(\mathbf{r}-\mathbf{r}^{'}) \propto \exp\{\frac{(\mathbf{r}-\mathbf{r}^{'})^T\mathbf{C}_{\mathbf{r}}(\mathbf{r}-\mathbf{r}^{'})}{-2h^2}\}$. Each kernel response is vectorised as $\mathbf{f}_{\mathbf{r}}$. Then vectors $\mathbf{f}_{\mathbf{r}}$ are collected in a local window ($3\times 3$) and in a center + surround window ($5\times 5$) both centered at $\mathbf{r}$ to form a feature matrix $\mathbf{F}^{(k)}_{|\mathbf{I}}$. The motivation for using such features stems from the fact that local regression kernels capture the underlying local structure of the data exceedingly well, even in the presence of significant distortions. Further they do not require explicit motion estimation.

As to object-based features, these are to be learned by specifically taking into account the classes of objects at hand. In the work presented here, the objects of interest are $\mathbf{O}=\{face, body\}$; thus, we compute face and person features by using the Haar/AdaBoost features exploited by the well-known Viola-Jones detector. This is a technical choice guided by computational efficiency issues; other choices [43], [37] would be equivalent from the modeling standpoint.

In order to be processed, features need to be spatially organized in feature maps. A feature map $\mathbf{X}$ is a topographically organized map that encodes the joint occurrence of a specific feature at a spatial location. It can be equivalently represented as a unique map encoding the presence of different object based features $\mathbf{F}^{(k)}_{f|\mathbf{O}}$ (e.g., face and body map), or a set of object-specific feature maps, i.e. $\mathbf{X}=\{\mathbf{X}_f\}$ (e.g., a face map, a body map, etc.). More precisely, referring to the $k$-th stream, $\mathbf{X}^{(k)}_f(t)$ is a matrix of binary RVs $x^{(k)}_f(\mathbf{r},t)$ denoting if feature $f$ is present or not present at location $\mathbf{r}$ at time $t$. Simply put, given $f$, $\mathbf{X}^{(k)}_f(t)$ is a map defining the spatial mask of $\mathbf{F}^{(k)}_{f|\mathbf{O}}$.

To support gaze-shift decisions, we define the RV $\mathbf{L}$ capturing the concept of priority map. Namely, for the $k$-th stream, denote $\mathbf{L}^{(k)}(t)$ the matrix of binary RVs $l^{(k)}(\mathbf{r},t)$ denoting if location $\mathbf{r}$ is to be considered relevant ($l^{(k)}(\mathbf{r},t)=1$) or not ($l^{(k)}(\mathbf{r},t)=0$) at time $t$. It is important to note that the term "relevant" is to be specified with respect to the kind of feature map used to infer a probability density function (pdf) over $\mathbf{L}^{(k)}$. For instance, if only bottom-up features are taken into account, then "relevant" boils down to "salient", and gaze shifts will be driven by the physical properties of the scene, such as motion, color, etc.

Eventually, in accordance with object-based attention approaches, we introduce *proto-objects* $\mathcal{O}^{(k)}(t)$ as the actual dynamic support for gaze orienting. Following Rensink [44], they are conceived as the dynamic interface between attentive and pre-attentive processing. Namely, a "quick and dirty" time-varying perception of the scene, from which a number of proto-objects is suitable to be glued in the percept of an object by the attentional process. Here, operatively, proto-objects are drawn from the priority map and each proto-object is used to sample *interest points* (IPs). The latter provide a sparse representation of a candidate objects to gaze at; meanwhile, the whole set of IPs sampled at time $t$ on video stream $k$ is used to compute the configurational complexity, say $\mathcal{C}^{(k)}(t)$, which is adopted as a prior quality index of the stream, in terms of foraging opportunities (potential preys within the patch).

More generally, whilst $\mathbf{X}$ and $\mathbf{L}$ can be conceived as perceptual memories, the dynamic ensemble of proto-objects is more similar to a working memory that allows attention to be temporarily focused on an internal representation [8].

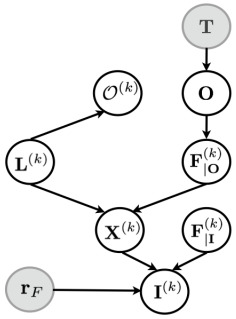


Fig. 2. The perception component (cfr. Fig. 1) as a Probabilistic Graphical Model. Graph nodes denote RVs and directed arcs encode conditional dependencies between RVs. Grey-shaded nodes stand for RVs whose value is given (current gaze position and task). Time index $t$ has been omitted for simplicity

Given the task $\mathbf{T}$ and the current gaze position $\mathbf{r}_F(t)$, perceptual inference relies upon the joint pdf $P(\mathbf{O},\mathbf{F}^{(k)}_{|\mathbf{O}}(t),\mathbf{L}^{(k)}(t),\mathcal{O}^{(k)}(t),\mathbf{X}^{(k)}(t),\mathbf{F}_{|\mathbf{I}},\mathbf{I}^{(k)}(t)|\mathbf{T},\mathbf{r}_F(t))$. The representation of such pdf can be given the form of the directed Probabilistic Graphical Model (PGM, [45]), say $\mathcal{G}$, presented in Fig. 2. The PGM structure captures the assumptions about the visual process previously discussed. For example, the assumption that given task $\mathbf{T}$, object class $\mathbf{O}$ is likely to occur, is represented through the dependence $\mathbf{T}\rightarrow\mathbf{O}$.

Stated technically, the $\mathcal{G}$ structure encodes the set $\mathcal{I}_\ell(\mathcal{G})$ of conditional independence assumptions over RVs (the local independencies, [45]) involved by the joint pdf. Then, the joint pdf factorizes according to $\mathcal{G}$ (cfr., Koller [45], Theorem 3.1):

$$P(\mathbf{O},\mathbf{F}^{(k)}_{|\mathbf{O}},\mathbf{L}^{(k)},\mathcal{O}^{(k)},\mathbf{X}^{(k)},\mathbf{F}^{(k)}_{|\mathbf{I}},\mathbf{I}^{(k)} \mid \mathbf{T},\mathbf{r}_F) = P(\mathbf{O}\mid\mathbf{T})P(\mathbf{F}^{(k)}_{|\mathbf{O}}\mid\mathbf{O})P(\mathbf{L}^{(k)})P(\mathcal{O}^{(k)}\mid\mathbf{L}^{(k)})\cdot P(\mathbf{X}^{(k)}\mid\mathbf{L}^{(k)},\mathbf{F}^{(k)}_{|\mathbf{O}})P(\mathbf{I}^{(k)}\mid\mathbf{F}^{(k)}_{|\mathbf{I}},\mathbf{X}^{(k)},\mathbf{r}_F) \quad (1)$$

(time index $t$ has been omitted for notational simplicity). The factorization specified in Eq. 1 makes explicit the local distributions (the set of independence assertions $\mathcal{I}(P)$ that hold in pdf $P$, $\mathcal{I}_\ell(\mathcal{G})\subseteq\mathcal{I}(P)$), and related inferences at the different levels of visual representation guiding gaze deployment.

*1) Object-based level:* $P(\mathbf{O}\mid\mathbf{T})$ is the multinomial distribution defining the prior on object classes under the given task, whose parameters can be easily estimated via Maximum-Likelihood (basically, object occurrence counting).

$P(\mathbf{F}^{(k)}_{|\mathbf{O}}(t)\mid\mathbf{O})$ represents the object-based feature likelihood. In current simulation, we use the Viola-Jones detector for faces and persons and convert the outcome to a probabilistic output (see [46], for a formal justification).

*2) Spatial-based level:* $P(\mathbf{L}^{(k)})$ denotes the prior probability of gazing at location $\mathbf{L}^{(k)}=\mathbf{r}^{(k)}$ of the scene. For example, specific pdfs can be learned to account for the gist of the scene [47] (given a urban scene, pedestrian are more likely to occur in the middle horizontal region) or specific spatial biases, e.g. the central fixation bias [40]. Here, we will not account for such tendencies, thus we assume a uniform prior. The factor $P(\mathcal{O}^{(k)}(t)|\mathbf{L}^{(k)}(t))$ is the proto-object likelihood given the priority map, which will be further detailed in Section V.

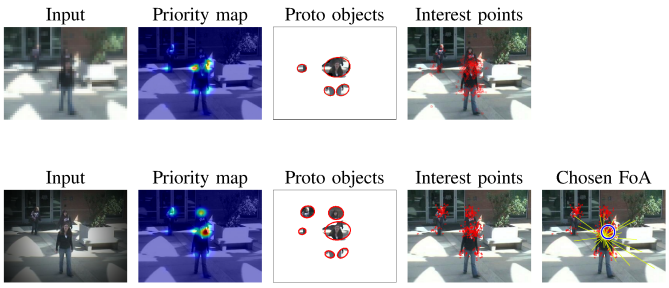


Fig. 3. The main perceptual representation levels involved by pre-attentive (top row) and attentive stages (bottom row). The input of the pre-attentive stage is the stream at low resolution. The priority map is visualized as a color map: reddish values specify most salient regions. Selected proto-objects are parametrised as ellipses. IPs sampled from proto-objects are displayed as red dots (cfr. Section IV). The input of the attentive stage is the foveated stream obtained by setting the initial FoA at the centre of the image. Candidates gaze shifts are displayed as yellow trajectories from the center of current FoA. The next FoA $\mathbf{r}_F(t+1)$ is chosen to maximise the expected reward (cfr. Section VI), and displayed as a white/blue circle.

*3) Feature map level:* $P(\mathbf{X}_f^{(k)}(t) \mid \mathbf{L}^{(k)}(t), \mathbf{F}_{|\mathbf{O}}^{(k)}(t))$ represents the likelihood of object-based feature $\mathbf{F}_{|\mathbf{O}} = \mathbf{f}_{|\mathbf{O}}$ to occur at location $\mathbf{L}^{(k)} = \mathbf{r}^{(k)}$. Following [43], when the feature is present at $\mathbf{r}^{(k)}$ we set $P(\mathbf{X}_f^{(k)}(t) = 1 \mid \mathbf{L}^{(k)}(t) = \mathbf{r}^{(k)}, \mathbf{F}_{|\mathbf{O}}^{(k)}(t) = 1)$ equal to a Gaussian $\mathcal{N}(\mathbf{r}^{(k)}, \sigma)$ centered at $\mathbf{r}^{(k)}$ ($\sigma = 1$), to activate nearby locations; otherwise, to a small value $P(\mathbf{X}_f^{(k)}(t) = 0 \mid \mathbf{L}^{(k)}(t) = \mathbf{r}^{(k)}, \mathbf{F}_{|\mathbf{O}}^{(k)}(t) = 0) = \epsilon$ ($\epsilon = 0.01$).

The factor $P(\mathbf{I}^{(k)}(t) \mid \mathbf{F}_{|\mathbf{I}}, \mathbf{X}_f^{(k)}(t), \mathbf{r}_F(t))$ is the feed-forward evidence obtained from low-level features $\mathbf{F}_{|\mathbf{I}}$ computed from frame $\mathbf{I}^{(k)}(t)$ as sensed when gaze is set at $\mathbf{r}_F(t)$. In the pre-attentive stage the position of gaze is not taken into account, and the input frame is a low-resolution representation of the original. In the attentive stage, $\mathbf{r}_F(t)$ is used to simulate foveation - accounting for the contrast sensitivity fall-off moving from the center of the retina, the fovea, to the periphery; thus, the input frame is a *foveated image* [37].

The feed-forward evidence is proportional to the output of low-level filters $f : \mathbf{I}(t) \rightarrow F_f(t)$. A variety of approaches can be used [26] from a simple normalization of filter outputs to more sophisticated Gaussian mixture modeling [47].

Here, based on the local regression kernel center/surround features, the evidence from a location $\mathbf{r}$ of the frame is computed as $P(\mathbf{I}^{(k)}(t) \mid \mathbf{x}_f^{(k)}(\mathbf{r}, t) = 1, \mathbf{F}_{|\mathbf{I}}, \mathbf{r}_F(t)) = \frac{1}{\sum_s} \exp\left(\frac{1-\rho(\mathbf{F}_{\mathbf{r}^{(k)},c}, \mathbf{F}_{\mathbf{r}^{(k)},s})}{\sigma^2}\right)$, where $\rho(\cdot) \in [-1, 1]$ is the matrix cosine similarity (see [42], for details) between center and surround feature matrices $\mathbf{F}_{\mathbf{r}^{(k)},c}$ and $\mathbf{F}_{\mathbf{r}^{(k)},s}$ computed at location $\mathbf{r}^{(k)}$ of the foveated frame.

Figure 3 illustrates main representations discussed above (spatio-temporal priority maps, proto-objects and IPs sampled from proto-objects)

### B. Action control

The model exploits a coarse-to-fine strategy. First, evaluation of stream "quality" is pre-attentively performed, resorting to the configurational complexity $\mathcal{C}^{(k)}(t)$ (cfr., Section IV). This stage corresponds to the *pre-attentive loop* briefly summarised in Algorithm 1. On this basis, the "best" quality

**Algorithm 1** Pre-attentive loop

$\{$*Parallel execution on all streams* $1, \cdots, K\}$
**Input:** $\{\mathbf{I}^{(k)}(t)\}_{k=1}^K$
Compute bottom-up feature $\mathbf{F}_{|\mathbf{I}}^{(k)}$ and weight the feature map $\mathbf{X}_f^{(k)}(t)$.
Sample the priority map $\mathbf{L}^{(k)}(t)$ conditioned on $\mathbf{X}^{(k)}(t)$.
Sample the potential object regions or proto-objects $\mathcal{O}^{(k)}(t)$ from $\mathbf{L}^{(k)}(t)$.
Based on available proto-objects $\mathcal{O}^{(k)}(t)$, sample IPs and compute the quality of the stream via complexity $\mathcal{C}^{(k)}(t)$

stream is selected (cfr., Section V), and the within-stream potential preys are attentively handled, in order to detect the actual targets $\mathbf{O}$ that are interesting under the given task $\mathbf{T}$ (cfr., Section VI). Intra-stream behavior thus boils down to an instance of the classic deployment of visual attention: spotting an object and keeping to it - via either fixational movements or smooth pursuit - or relocating to another region (saccade) [39]. Thus, intra-stream behavior does not reduce to tracking, which is to be considered solely the computational realization of visual smooth pursuit. The attentive loop is summarised in Algorithm 2.

**Algorithm 2** Attentive loop

**Input:** $\{\mathbf{I}^{(k)}(t)\}_{k=1}^K$, $\{\mathcal{C}^{(k)}(t)\}_{k=1}^K$
$\{$*Patch choice*$\}$
Based on the complexities $\{\mathcal{C}^{(k)}(t)\}_{k=1}^K$, sample the video stream $\widehat{k}$ to be analyzed.
Execute the *between-stream gaze shift* $\mathbf{r}_F(t-1) \mapsto \mathbf{r}_F(t)$ at the center of current frame of stream $\widehat{k}$ and set the current FoA;
**repeat**
  Compute bottom-up and top-down features $\{\mathbf{F}_{|\mathbf{I}}^{(\widehat{k})}(t), \mathbf{F}_{|\mathbf{O}}^{(\widehat{k})}(t)\}$ and sample the feature map $\mathbf{X}^{(\widehat{k})}(t)$, based on $\mathbf{F}_{|\mathbf{O}}^{(\widehat{k})}(t)$
  Sample the priority map $\mathbf{L}^{(k)}(t)$ conditioned on $\mathbf{X}^{(k)}(t)$.
  Sample proto-objects $\mathcal{O}^{(\widehat{k})}(t)$ from $\mathbf{L}^{(k)}(t)$;
  Based on $\mathcal{O}^{(k)}(t)$, sample IPs and compute the quality of the stream via complexity $\mathcal{C}^{(\widehat{k})}(t)$
  $\{$*Prey handling*$\}$
  Execute the *within-stream gaze shift* $\mathbf{r}_F(t) \mapsto \mathbf{r}_F(t+1)$ in order to maximize the expected reward with respect to the IP value and analyze the current FoA.
**until** giving-up condition is met
$\{$*Patch leave*$\}$

Clearly, since the number of targets is a priori unknown (partial information condition), efficient search requires tailoring a stopping decision to target handling within the stream. From a foraging standpoint, the decision to leave should also depend

on future prospects for food on the current patch, which in turn depends on posterior information about this patch. This issue is addressed in the framework of optimal Bayesian foraging [48], [49] (cfr., Section VII).

## IV. Pre-attentive sensing

The goal of this stage is to infer a proto-object representation of all the $K$ patches within the spatial landscape (Fig. 3, top row). To this end, the posterior $P(\mathbf{L}^{(k)}(t) \mid \mathbf{I}^{(k)}(t)) \approx P(\mathbf{L}^{(k)}(t), \mathbf{I}^{(k)}(t))$ is calculated from the joint pdf. In the derivations that follows we omit the time index $t$ for notational simplicity.

Rewrite the joint pdf factorization in Eq. 1 under the assumption of object-based feature independence, i.e., $\prod_{f,f'} P(\mathbf{O}, \mathbf{F}^{(k)}_{f|\mathbf{O}}, \mathbf{L}^{(k)}, \mathcal{O}^{(k)}, \mathbf{X}^{(k)}_f, \mathbf{F}_{f'|\mathbf{I}}, \mathbf{I}^{(k)} | \mathbf{T}, \mathbf{r}_F)$. Then $P(\mathbf{L}^{(k)}, \mathbf{I}^{(k)})$ is obtained by marginalizing over RVs $\mathbf{X}^{(k)}_f, \mathbf{F}^{(k)}_{f|\mathbf{O}}, \mathbf{F}^{(k)}_{f'|\mathbf{I}}, \mathbf{O}$ and $\mathcal{O}^{(k)}$. Use the following: $\sum_{\mathcal{O}} P(\mathcal{O}^{(k)}|\mathbf{L}^{(k)}) = 1$ by definition; $P(\mathbf{L}^{(k)}) = Unif$ by assumption; $\sum_{\mathcal{O}} P(\mathbf{F}^{(k)}_{f|\mathbf{O}}|\mathbf{O}) P(\mathbf{O}|\mathbf{T}) = P(\mathbf{F}^{(k)}_{f|\mathbf{O}}|\mathbf{T}) = P(\mathbf{F}^{(k)}_{f|\mathbf{O}})$ by local conditional independence in $\mathcal{G}$. Thus:

$$P(\mathbf{L}^{(k)} \mid \mathbf{I}^{(k)}) \approx \prod_{f,f'} \sum_{\mathbf{X}^{(k)}_f, \mathbf{F}^{(k)}_{f|\mathbf{O}}, \mathbf{F}^{(k)}_{f'|\mathbf{I}}} P(\mathbf{X}^{(k)}_f \mid \mathbf{L}^{(k)}(t), \mathbf{F}^{(k)}_{f|\mathbf{O}}) \, P(\mathbf{I}^{(k)} \mid \mathbf{F}^{(k)}_{f'|\mathbf{I}}, \mathbf{X}^{(k)}_f, \mathbf{r}_F) P(\mathbf{F}^{(k)}_{f|\mathbf{O}}). \quad (2)$$

The term $P(\mathbf{F}^{(k)}_{f|\mathbf{O}})$ is a prior "tuning" the preference for specific object-based features. In the pre-attentive stage we assume a uniform prior, i.e. $P(\mathbf{F}^{(k)}_{f|\mathbf{O}}) = Unif.$, and restrict to feed-forward features $\mathbf{F}^{(k)}{}_{|\mathbf{I}}$. Then, Eq. 2 boils down to the probabilistic form of a classic feed-forward saliency map (see Fig. 3), namely,

$$P(\mathbf{L}^{(k)} \mid \mathbf{I}^{(k)}) \approx \prod_{f,f'} \sum_{\mathbf{X}^{(k)}_f, \mathbf{F}^{(k)}_{f|\mathbf{O}}, \mathbf{F}^{(k)}_{f'|\mathbf{I}}} P(\mathbf{X}^{(k)}_f \mid \mathbf{L}^{(k)}, \mathbf{F}^{(k)}_{f|\mathbf{O}}) \cdot P(\mathbf{I}^{(k)} \mid \mathbf{F}^{(k)}_{f'|\mathbf{I}}, \mathbf{X}^{(k)}_f, \mathbf{r}_F), \quad (3)$$

where the likelihood $P(\mathbf{X}^{(k)}_f | \mathbf{L}^{(k)}, \mathbf{F}^{(k)}_{f|\mathbf{O}})$ is modulated by bottom-up feature likelihood $P(\mathbf{I}^{(k)} | \mathbf{F}^{(k)}_{f'|\mathbf{I}}, \mathbf{X}^{(k)}_f, \mathbf{r}_F)$.

Given the priority map, a set $\mathcal{O}^{(k)}(t) = \{\mathcal{O}^{(k)}_p(t)\}_{p=1}^{N_P}$ of $N_P$ proto-objects or candidate preys can be sampled from it. Following [39], we exploit a sparse representation of proto-objects. These are conceived in terms of "potential bites", namely interest points sampled from the proto-object. At any given time $t$, each proto-object is characterised by different shape and location, i.e., $\mathcal{O}^{(k)}_p(t) = (O^{(k)}_p(t), \Theta_p(t))$. Here $O^{(k)}_p(t) = \{\mathbf{r}^{(k)}_{i,p}\}_{i=1}^{N_{i,p}}$ is the sparse representation of proto-object $p$ as the cluster of $N_{i,p}$ IPs sampled from it; $\Theta^{(k)}_p(t)$ is a parametric description of a proto-object, $\Theta^{(k)}_p(t) = (\mathcal{M}^{(k)}_p(t), \theta^{(k)}_p)$.

The set $\mathcal{M}^{(k)}_p(t) = \{m^{(k)}_p(\mathbf{r}, t)\}_{\mathbf{r}\in L}$ stands for a map of binary RVs indicating at time $t$ the presence or absence of proto-object $p$, and the overall map of proto-objects is given by $\mathcal{M}^{(k)}(t) = \bigcup_{p=1}^{N_p} \mathcal{M}^{(k)}_p(t)$. Location and shape of the proto-object are parametrized via $\theta^{(k)}_p$. Assume independent proto-objects:

$$\mathcal{M}^{(k)}(t) \sim P(\mathcal{M}^{(k)}(t) | \mathbf{L}^{(k)}(t)), \quad (4)$$

and for $p = 1, \cdots, N_P$

$$\theta^{(k)}_p(t) \sim P(\theta^{(k)}_p(t) | \mathcal{M}^{(k)}_p(t) = 1, \mathbf{L}^{(k)}(t)), \quad (5)$$

$$O^{(k)}_p(t) \sim P(O^{(k)}_p(t) | \theta^{(k)}_p(t), \mathcal{M}^{(k)}_p(t) = 1, \mathbf{L}^{(k)}(t)). \quad (6)$$

The first step (Eq. 4) samples the proto-object map from the landscape. The second (Eq. 5) samples proto-object parameters $\theta(t)^{(k)}_p = (\mu^{(k)}_p(t), \Sigma^{(k)}_p(t)))$.

Here, $\mathcal{M}^{(k)}(t)$ is drawn from the priority map by deriving a preliminary binary map $\widetilde{\mathcal{M}}^{(k)}(t) = \widehat{m}^{(k)}(\mathbf{r}, t)\}_{\mathbf{r}\in L}$, such that $\widehat{m}^{(k)}(\mathbf{r}, t) = 1$ if $P(\mathbf{L}^{(k)}(t) | \mathbf{I}^{(k)}(t)) > T_M$, and $\widehat{m}^{(k)}(\mathbf{r}, t) = 0$ otherwise. The threshold $T_M$ is adaptively set so as to achieve $95\%$ significance level in deciding whether the given priority values are in the extreme tails of the pdf. The procedure is based on the assumption that an informative proto-object is a relatively rare region and thus results in values which are in the tails of $P(\mathbf{L}^{(k)}(t) | \mathbf{I}^{(k)}(t))$. Then, following [50], $\mathcal{M}^{(k)}(t) = \{\mathcal{M}^{(k)}_p(t)\}_{p=1}^{N_P}$ is obtained as $\mathcal{M}^{(k)}_p(t) = \{m^{(k)}_p(\mathbf{r}, t) | \ell(B, \mathbf{r}, t) = p\}_{\mathbf{r}\in L}$, where the function $\ell$ labels $\widetilde{\mathcal{M}}(t)$ around $\mathbf{r}$.

We set the maximum number of proto-object to $N_P = 15$ to retain the most important ones.

As to Eq. 5, the proto-object map provides the necessary spatial support for a 2D ellipse maximum-likelihood approximation of each proto-object, whose location and shape are parametrized as $\theta^{(k)}_p = (\mu^{(k)}_p, \Sigma^{(k)}_p)$ for $p = 1, \cdots, N_p$ (see [39] for a formal justification).

In the third step (Eq. 6), the procedure generates clusters of IPs, one cluster for each proto-object $p$ (see Fig. 3). By assuming a Gaussian distribution centered on the proto-object - thus with mean $\mu^{(k)}_p$ and covariance matrix $\Sigma^{(k)}_p$ given by the axes parameters of the 2D ellipse fitting the proto-object shape -, Eq. (6) can be further specified as [39]:

$$\mathbf{r}^{(k)}_{i,p} \sim \mathcal{N}(\mathbf{r}^{(k)}_p; \mu^{(k)}_p(t), \Sigma^{(k)}_p(t)), i = 1, \cdots, N_{i,p}. \quad (7)$$

We set $N_s = 50$ the maximum number of IPs and for each proto-object $p$, we sample $\{\mathbf{r}^{(k)}_{i,p}\}_{i=1}^{N_{i,p}}$ from a Gaussian centered on the proto-object as in (7). The number of IPs per proto-object is estimated as $N_{i,p} = \lceil N_s \times \frac{A_p}{\sum_p A_p} \rceil$, $A_p = \pi \sigma_{x,p} \sigma_{y,p}$ being the size (area) of proto-object $p$. Eventually, the set of all IPs characterising the pre-attentively perceived proto-object can be obtained as $O(t) = \bigcup_{p=1}^{N_p} \{\mathbf{r}^{(k)}_{i,p}(t)\}_{i=1}^{N_{i,p}}$.

## V. Stream selection

Streams vary in the number of objects they contain and maybe other characteristics such as the ease with which individual items are found. We assume that in the pre-attentive stage, the choice of the observer to spot a stream, is drawn on the basis of some global index of interest characterizing each stream in the visual landscape. In ecological modelling for instance, one such index is the landscape entropy determined by dispersion/concentration of preys [1].

Here, generalizing these assumptions, we introduce the time-varying configurational complexity $\mathcal{C}^{(k)}(t)$ of the $k$-th stream. Intuitively, by considering each stream a dynamic system, we resort to the general principle that complex systems are neither completely random neither perfectly ordered and complexity should reach its maximum at a level of randomness away from these extremes [51]. For instance, a crowded scene with many pedestrians moving represents a disordered system (high entropy, low order) as opposed to a scene where no activities take place (low entropy, high order). The highest complexity is thus reached when specific activities occur: e.g., a group of people meeting. To formalize the relationship between stream complexity and stream selection we proceed as follows. Given $\mathcal{C}^{(k)}(t), k = 1, \cdots, K$, the choice of the $k$-th stream is obtained by sampling from the categorical distribution

$$k \sim \prod_{k=1}^{K} \left[ P(\mathcal{C}^{(k)}(t)) \right]^{k}, \tag{8}$$

with

$$P(\mathcal{C}^{(k)}(t)) = \frac{\mathcal{C}^{(k)}(t))}{\sum_{k=1}^{K} \mathcal{C}^{(k)}(t))}. \tag{9}$$

Keeping to [51], complexity $\mathcal{C}^{(k)}(t)$ is defined in terms of order/disorder of the system,

$$\mathcal{C}^{(k)}(t) = \Delta^{(k)}(t) \cdot \Omega^{(k)}(t), \tag{10}$$

where $\Delta^{(k)} \equiv H^{(k)}/H^{(k)}_{sup}$ is the disorder parameter, $\Omega^{(k)} = 1 - \Delta^{(k)}$ is the order parameter, and $H^{(k)}$ the Boltzmann-Gibbs-Shannon (BGS) entropy with $H^{(k)}_{sup}$ its supremum. $H^{(k)}$ and $H^{(k)}_{sup}$ are calculated as follows.

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

Thus, $H^{(k)}(t) = -k_B \sum_{c=1}^{N_w} P^{(k)}(c,t) \log P^{(k)}(c,t)$, and (10) can be easily computed. Since dealing with a fictitious thermodynamical system, we set Boltzmann's constant $k_B = 1$. The supremum of $H^{(k)}(t)$ is $H_{sup} = \log N_w$ and it is associated to a completely unconstrained process, that is a process where $H^{(k)}(t) = const$, since with reflecting boundary conditions the asymptotic distribution is uniform.

When stream $k$ is chosen at time $t-1$, attention is deployed to the stream via the gaze shift $\mathbf{r}_F(t-1) \rightarrow \mathbf{r}_F(t)$, and the "entering time" $t_{in} = t$ is set.

## VI. Attentive stream handling

When gaze is deployed to the $k$-th stream, the $\mathbf{r}_F(t_{in})$ is positioned at the centre of the frame, and foveation is simulated by blurring $\mathbf{I}^{(k)}(t_{in})$ through an isotropic Gaussian function centered at $\mathbf{r}_F(t_{in})$, whose variance is taken as the radius of a FoA, $\sigma = |FOA|$. This is approximately given by $1/8 \min[width, height]$, where $width \times height = |\mathcal{D}|$, $|\mathcal{D}|$ being the dimension of the frame support $\mathcal{D}$. This way we obtain the foveated image, which provides the input for the next processing steps. The foveation process is updated for every gaze shift within the patch that involves a large relocation (saccade), but not during small relocations, i.e. fixational or pursuit eye movements. At this stage, differently from pre-attentive analysis, the observer exploits the full priority posterior as formulated in Eq. 2, rather than the reduced form specified in Eq. 3. In other terms, the object-based feature likelihood, $P(\mathbf{F}^{(k)}_{|\mathbf{O}}|\mathbf{O})$, is taken into account.

Object search is performed by sampling, from current location $\mathbf{r}_F$, a set of candidate gaze shifts $\mathbf{r}_F(t) \rightarrow \mathbf{r}^{(k)}_{new}(t+1)$ (cfr. Fig.3, bottom-right picture). In simulation, candidate point sampling is performed as in [39]. In a nutshell, $\mathbf{r}^{(k)}_{new}(t+1)$ are sampled via a Langevin-type stochastic differential equation, where the drift component is a function of IPs' configuration, and the stochastic component is sampled from the Lévy $\alpha$-stable distribution. The latter accounts for prior oculomotor biases on gaze shifts. We use different $\alpha$-stable parameters for the different types of gaze shifts - fixational, pursuit and saccadic shifts -, that have been learned from eye-tracking experiments of human subjects observing videos under the same task considered here. The time-varying choice of the family of parameters is conditioned on the current complexity index $\mathcal{C}^{(k)}(t)$ ([39] for details).

Denote $R^{(k)}$ the reward consequent on a gaze shift. Then, next location is chosen to maximize the expected reward:

$$\mathbf{r}_F(t+1) = \arg\max_{\mathbf{r}^{(k)}_{new}} E\left[ R^{(k)}_{\mathbf{r}^{(k)}_{new}} \right]. \tag{11}$$

The expected reward is computed with reference to the value of proto-objects available within the stream,

$$E\left[ R^{(k)}_{\mathbf{r}^{(k)}_{new}} \right] = \sum_{p \in \mathcal{I}^{(k)}_V} Val(\mathcal{O}^{(k)}_p(t)) P(\mathcal{O}^{(k)}_p(t) | \mathbf{r}^{(k)}_{new}(t+1), \mathbf{T}). \tag{12}$$

Here $Val$ is the average value of proto-object $\mathcal{O}_p(t)$ with respect to the posterior $P(\mathbf{L}^{(k)}(t)|\mathbf{I}^{(k)}(t))$, which, by using samples generated via Eq. 7, can be simply evaluated as

$$Val(\mathcal{O}^{(k)}_p(t)) \simeq \sum_{i \in \mathcal{I}_p} P(L^{(k)}_i(t)|\mathbf{I}^{(k)}(t)). \tag{13}$$

The observer samples $N_{new}$ candidate gaze shifts. Using Eqs. 7 and 13, Eq. 12 can be written as

$$E\left[ R^{(k)}_{\mathbf{r}^{(k)}_{new}} \right] = \sum_{p \in \mathcal{I}^{(k)}_V} \sum_{i \in \mathcal{I}_p} Val(\mathbf{r}^{(k)}_{i,p}(t)) \mathcal{N}(\mathbf{r}^{(k)}_{i,p}(t) | \mathbf{r}^{(k)}_{new}(t+1), \Sigma_s), \tag{14}$$

where $\Sigma_s$ defines the region around $\mathbf{r}^{(k)}_{new}(t+1)$. In foraging terms, Eq. 12 formalises the expected reward of gaining valuable bites of food (IPs) in the neighbourhood of the candidate shift $\mathbf{r}_{new}$.

Note that effective reward $R^{(k)}(t)$ is gained by the observer only if the gaze shift is deployed to a point $\mathbf{r}$ that sets a FoA overlapping an object of interest for the task (in the simulation, for simplicity, $R^{(k)}(t) = 1$ when a face or a body is detected, and 0 in other cases). Thus, as the observer attentively explores the stream, he updates his estimate of stream quality in terms

of accumulated rewards, which will provide the underlying support for the stream giving-up strategy.

A final remark concerns the number of objects that can be detected within the stream. Attentive analysis is sequential by definition. In principle, all relevant objects in the scene can be eventually scrutinized, provided that enough time is granted to the observer. For instance, as to detection performance, current implementation of the model exploits *adaboost* face and body detectors that have been trained on a much larger dataset than original Viola-Jones detectors, leading to about $90\%$ detection accuracy (considering a minimal detectable region of $40 \times 40$ pixel area). But cogently, the actual number of scrutinized objects is the result of observer's trade-off between the quality of the visited stream and the potential quality of the other $K-1$ streams. Namely, it depends on the stream giving-up time as dynamically determined by the Bayesian strategy.

## VII. The Bayesian giving-up strategy

In this Section we consider the core problem of switching from one stream to another. In foraging theory this issue is addressed as "How long should a forager persevere in a patch?". Two approaches can be pursued: i) patch-based or global/distal models; ii) prey-based or local/proximal models. These are, for historical reasons, subject to separate analyses and modeling [1]. The Bayesian strategy we propose here aims at filling such gap.

### A. Global models. Charnov's Marginal Value Theorem

In the scenario envisaged by Charnov [5] the landscape is composed of food patches that deliver food rewards as a smooth decreasing flow. Briefly, Charnov's MVT states that a patch leave decision should be taken when the expected current rate of information return falls below the mean rate that can be gained from other patches. MVT considers food intake as a continuous deterministic process where foragers assess patch profitability by the instantaneous net energy intake rate. In its original formulation, it provides the optimal solution to the problem, although only once the prey distribution has already been learnt; it assumes omniscient foragers (i.e. with a full knowledge of preys and patch distribution). The model is purely functional, nevertheless it is important for generating two testable qualitative predictions [52]: 1) patch time should increase with prey density in the patch; 2) patch times should increase with increasing average travel time in the habitat and should decrease with increasing average host density in the patches.

### B. Local models

The MVT and its stochastic generalization do not take into account the behavioral *proximate* mechanisms used by foragers to control patch time or to obtain information about prey distribution [52]. Such a representation of intake dynamics is inadequate to account for the real search/capture processes occurring within the patch. These, in most cases, are discrete and stochastic events in nature. For instance, Wolfe [4] has examined human foraging in a visual search context, showing that departures from MVT emerge when patch quality varies and when visual information is degraded.

Experience on a patch, in terms of cumulative reward, gives information on current patch type and on future rewards. A good policy should make use of this information and vary the giving-up time with experience. In this perspective, as an alternative to MVT, local models, e.g., Waage's [38], assume that the motivation of a forager to remain and search on a particular patch would be linearly correlated with host density. As long as this "responsiveness" is above a given (local) threshold, the forager does not leave the patch [38]. As a consequence, the total time spent within the patch, say $\Delta_w^{(k)}$, eventually depends on the experience of the animal within that patch.

### C. Distal and proximal strategies in an uncertain world

To deal with uncertainty [48], [49], a forager should persevere in a patch as long as the probability of the next observation being successful is greater than the probability of the first observation in one among the $K-1$ patches being successful, taking into account the time it takes to make those observations.

Recall that complexity $\mathcal{C}^{(k)}(t)$ is used as a pre-attentive stochastic proxy of the likelihood that the $k$-th stream yields a reward $R^{(k)}(t)$ to the observer. Thus, $P(\mathcal{C}^{(k)}(t))$ defined in Eq. 9 stands for the prior probability of objects being primed for patch $k$ (in OFT, the base rate [1]).

A common detection or gain function, that is the probability of reinforcement vs. time, is an exponential distribution of times to detection [49], and can be defined in terms of the conditional probability of gaining a reward in stream $k$ by time $t$, given that it has been primed via complexity $\mathcal{C}^{(k)}(t)$:

$$P(R^{(k)}(t) \mid \mathcal{C}^{(k)}(t)) = 1 - \exp(-\lambda t) \tag{15}$$

where $\lambda$ is the detection rate. Then, by generalising the two patch analysis discussed in [49], the following holds.

*Proposition 7.1:* Denote $\langle \mathcal{C}^{(k)}(t) \rangle_{i \neq k}$ the average complexity of the $K-1$ streams other than $k$. Under the hypothesis that at $t=0$, $C^{(k)}(0) > \langle \mathcal{C}^{(k)}(t) \rangle_{i \neq k}$, leaving the $k$-th stream when

$$\mathcal{C}^{(k)}(t) \exp(-\lambda t) = \langle \mathcal{C}^{(k)}(t) \rangle_{i \neq k}, t > 0, \tag{16}$$

defines an optimal Bayesian strategy for the observer.

*Proof:* See Appendix A. ∎

The strategy summarised via Eq. 16 can be considered as a Bayesian version of the MVT-based strategy [5]. In order to reconcile the *distal* functional constraint formalised through Eq. 16, with the behavioral proximate mechanisms used by foragers within the patch, we put a prior distribution on the $\lambda$ parameter of the exponential distribution in the form of a Gamma distribution, i.e., $Gamma(\lambda; \nu^{(k)}, \Delta^{(k)})$, where $\nu^{(k)}, \Delta^{(k)}$ are now hyper-parameters governing the distribution of the $\lambda$ parameter. Assume that when the observer selects the stream, the initial prior is $Gamma(\lambda; \nu_0^{(k)}, \Delta_0^{(k)})$.

The hyper-parameters $\nu_0^{(k)}, \Delta_0^{(k)}$ represent initial values of expected rewards and "capture" time, respectively, thus stand for "a priori" estimates of stream profitability. For $t > t_{in}$,

the posterior over $\lambda$ can be computed via Bayes' rule as $Gamma(\lambda;\nu^{(k)},\Delta^{(k)}) \propto \exp(-\lambda t)Gamma(\lambda;\nu_0^{(k)},\Delta_0^{(k)})$. Since the Gamma distribution is a conjugate prior, the Bayesian update only calls for the determination of the hyper-parameter update

$$\nu^{(k)} = \nu_0^{(k)} + n, \qquad \Delta^{(k)} = \Delta_0^{(k)} + \sum_{i=1}^{n} \Delta(t_n), \qquad (17)$$

$n$ being the number of handled objects, that is the number of rewards effectively gained up to current time, i.e., $\sum_{t=t_{in}}^{t'} R^{(k)}(t)$, and $\Delta(t_n)$ the interval of time spent on the $n$-th proto-objects. The latter, in general, can be further decomposed as $\Delta(t_n) = TD_n + TH_n$, where $TD_n$ and $TH_n$ denote the time to spot and handle the $n$-th proto-object, respectively. Clearly, time $TD_n$ elapses for any proto-object within the stream, whilst $TH_n$ is only taken into account when the object has been detected as such (e.g., a moving proto-object as a pedestrian) and actual object handling occurs (e.g., tracking the pedestrian), otherwise $TH_n = 0$. In the experimental analyses we will assume, for generality, $TD_n = \delta_D \phi(|Proto|)$ and $TH_n = \delta_H \phi(|Object|)$, where $\delta_D$ and $\delta_H$ are times to process elements (pixels, super-pixels, point representation or parts) defining the prey, which depend on the specific algorithm adopted; $\phi(|\cdot|)$ is a function (linear, quadratic, etc) of the dimension of the processed item.

Eventually, when hyper-parameters have been computed (Eq. 17), a suitable value for $\lambda$ can be obtained as the expected value $\overline{\lambda} = E_{Gamma}[\lambda] = \frac{\nu^{(k)}}{\Delta^{(k)}}$. As a consequence, the total within-stream time $\Delta_w^{(k)}$ depends on the experience of the observer within that stream

Here, this proximal mechanism is formally related to the distal global quality of all streams, via the condition specified through Eq. 16 so that the decision threshold is dynamically modulated by the pre-attentive observer's perception across streams. As a result, even though on a short-time scale the observer might experience local motivational increments due to rewards, on a longer time scale the motivation to stay within the current stream will progressively decrease.

## VIII. Experimental work

### A. Dataset

We used a portion of the the UCR Videoweb Activities Dataset [53], a publicly available dataset containing data recorded from multiple outdoor wireless cameras. The dataset contains $4$ days of recording and several scenes for each day, about $2.5$ hours of video displaying dozens of activities along with annotation. For the first three days, each scene is composed of a collection of human activities and motions which forms a continuous storyline.

The dataset is designed for evaluating the performance of human-activity recognition algorithms, and it features multiple human activities viewed from multiple cameras located asymmetrically with overlapping and non-overlapping views, with varying degrees of illumination and lighting conditions. This amounts to a large variety of simple actions such as walking, running, and waving.

We experimented on three different scenes recorded in three different days. Here we present results obtained from scene 1, recorded in the second day (eight camera recordings). Results from the other scenes are reported as Supplementary Material. The scene contains the streams identified by the following ids: $cam16$, $cam17$, $cam20$, $cam21$, $cam27$, $cam31$, $cam36$, $cam37$. Each video is at 30 fps and cameras are not time-synchronized. We synchronized video streams by applying the following shifts between cameras: $[cam16:291, cam17:191, cam20:0, cam21:0, cam27:389, cam31:241, cam36:0, cam37:373]$. Cameras $cam20$, $cam21$ and $cam36$ can be used as time reference. Since the video of the camera $cam21$ is the shortest ($\approx 8000$ frames), the analyzes presented in the following consider the frames between 1 and 8000.

Annotated activities are: *argue within two feet*, *pickup object*, *raised arms*, *reading book*, *running*, *sit cross legged*, *sit on bench*, *spin while talking*, *stand up*, *talk on phone*, *text on phone*. All are performed by humans.

As previously discussed, we are not concerned with action or activity recognition. Nevertheless, the dataset provides a suitable benchmark. The baseline aim of the model is to dynamically set the FoA on the most informative subsets of the video streams in order to capture atomic events that are at the core of the activities actually recorded. In this perspective, the virtual forager operates under the task "pay attention to people within the scene", so that the classes of objects of interest are represented by faces and human bodies. The output collection of subsets from all streams can eventually be evaluated in terms of the retrieved activities marked in the ground-truth.

### B. Experimental evaluation

Evaluation of results should consider the two dimensions of i) visual representation and ii) giving-up strategy. For instance, it is not straightforwardly granted that a pre-attentive representation for choosing the patch/video might perform better (beyond computational efficiency considerations) than an attentive representation, where all objects of interest are detected before selecting the video stream.

As to the giving-up time choice, any strategy should in principle perform better than random choice. Again, this should not be given for granted, since in a complex scenario a bias-free, random allocation could perform better than expected. Further, a pure Charnov-based strategy, or a deterministic one, e.g. [36], could offer a reasonable solution. Under this rationale, evaluation takes into account the following analyses.

*1) Representations of visual information:* Aside from the basic priority map representation (denoted M in the remainder of this Section), which is exploited by our model (Eqs. 3 and 2 for the pre-attentive and attentive stages, respectively), the following alternatives have been considered.

- *Static* (denoted S): the baseline salience computation by Itti *et al.* [54]. The method combines orientation, intensity and color contrast features in a purely bottom-up scheme. The frame-based saliency map is converted in a probability map (as in [37]) so to implement the bottom-up priority map (Eq. 3). Attentive exploration is driven

by bottom-up information and object-based likelihood is kept uniform when computing Eq. 2.

- *Static combined with Change Detection and Face/Body Detection* (S+CD+FB): this representation has been used in [36]. It draws on the Bayesian integration of top-down / bottom-up information as described in [37]. Novelties are computed by detecting changes between two subsequent frames at a lower spatial resolution [36]. In our setting, it amounts to assume that the observer has the capability of detecting objects before selecting the stream; namely, it boils down to directly compute Eq. 2.
- *Proposed model with early prey detection* (M+): akin to the S+CD+FB scheme, the full priority probability (Eq. 2) is exploited before stream selection, instead of the bottom-up priority (Eq. 3).

Clearly, there are differences between adopting one representation or the other. These can be readily appreciated by analyzing behavior over time of stream complexities $\mathcal{C}^{(k)}, k = 1, \cdots, K$ obtained by adopting the above representations. One stream is hardly distinguishable from another when using the S and the S+CD+FB representations; by contrast, higher discriminability is apparent for the M and M+ settings (cfr. Fig 12 and 13, Supplementary Material). Yet, most interesting here is to consider representational performance as related to foraging strategy.

*2) Foraging strategies:* As to stream giving-up, we compare the following strategies.

*a) Deterministic:* The simplest strategy [1]. A camera switch is triggered after a fixed time $\Delta_w > 0$. Higher values of within-stream time $\Delta_w$ entail a low number of switches.

*b) Random:* This strategy triggers a camera switch after a random time $\Delta_w$. In this case $\Delta_w$ is a RV drawn from a uniform pdf $Unif(0, b_w)$, where $b_w$ is a suitable parameter.

*c) Charnov:* We adapted the solution to Charnov's MVT [5] by Lundberg et *al* [55]. If the observer chooses stream $k$ at time $t_{in}$, the optimal stream residence time is defined as $\Delta_w^{(k)} = \mathcal{C}^{(k)}(t_{in}) \cdot \sqrt{\frac{t_b}{\langle \mathcal{C}(t_{in}) \rangle \cdot \delta}}$, where $\mathcal{C}^{(k)}(t_{in})$ is the resource level in stream $k$ (here assessed in terms of complexity) at entering time $t_{in}$, $\langle \mathcal{C}(t_{in}) \rangle$ is the average resource level across streams, $\delta$ is a parameter determining the initial slope of the gain function in the stream, and $t_b$ is the average switching (travelling) time between two video streams. By assuming constant traveling time ($t_b = 1$), the only parameter to determine is the slope $\delta > 0$. Note that, when $\mathcal{C}^{(k)}(t_{in}) > \langle \mathcal{C}(t_{in}) \rangle$, higher values of $\mathcal{C}^{(k)}(t_{in})$ entail higher values of $\Delta_w^{(k)}$.

### C. Evaluation measures

The definition of measures that capture the subtleties of activity dynamics across multiple cameras is not straightforward. One has to account for the overall distribution of activities with respect to the different streams. Meanwhile, each stream should be characterized in terms of the activities as evolving in time.

As to the first issue, consider the joint probability $P(k, e)$ where $k$ can now be considered as a discrete RV indexing the streams and $e$ is a discrete RV indexing the given activity set (*argue within two feet*, etc.). Such joint distribution can be empirically estimated as $N(k = i, e = j) / \sum_k \sum_e N(k, e)$, where $N(k = i, e = j)$ denotes the number of frames of stream $i$ that displays activity $j$. In Fig. 4, $P(k, e)$ is rendered as a 2D Hinton diagram. The joint distribution is suitable to provide two essential pieces of information.

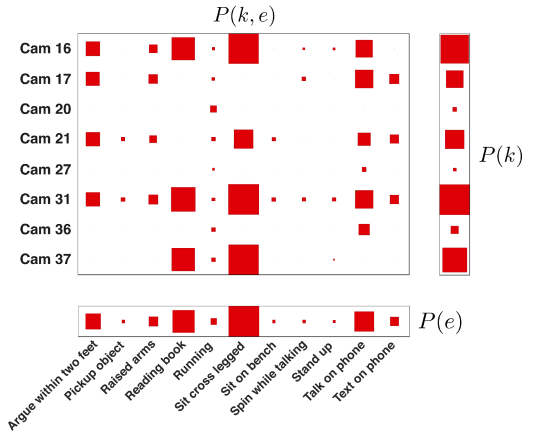


Fig. 4. A probabilistic glance at the information content of multiple video streams: the distribution of activities within each stream of the dataset represented in terms of joint and marginal probability distributions $P(k, e)$, $P(k)$ and $P(e)$, respectively. Distributions are visualised as Hinton diagrams, i.e., the square size is proportional to the probability value. The plots for the marginal distributions have different scales from those for the joint distribution (on the same scale, the marginals would look larger as they sum all of the mass from one direction).

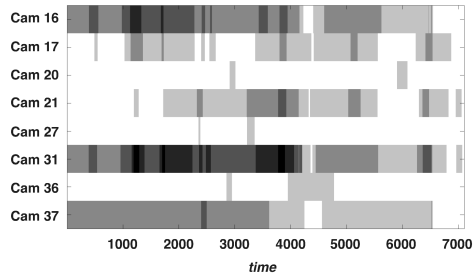


Fig. 5. Distribution of activities across cameras and time. A darker color indicates that several activities co-occur at the same time while the white color indicates the total absence of activities.

On the one hand, the marginalization of $P(k, e)$ over $e$, i.e., $P(k) = \sum_e P(k, e)$, offers an insight into the relevance of each stream - in terms of the marginal likelihood $P(k)$ - to the job of collating informative stream subsets (cfr. Fig. 4). Intuitively, we expect the corresponding marginal computed after stream processing, say $\widetilde{P}(k)$, to be a sparse summarisation of $P(k)$. Yet, it should account for the original representational relevance of the streams (cfr. Fig. 6). This can be further understood by displaying the distribution of activities across cameras and time as in Fig. 5 (a darker color indicates several activities co-occurring at the same time). By comparing with the Hinton diagram of $P(k)$ in Fig. 4, it is readily seen that some cameras capture a large amount of activities - for instance $cam16$ and $cam31$ -, whilst other cameras, e.g., $cam20$ and $cam27$, feature few activities. At the same time, the information displayed by subsets of one stream can be considered as redundant with respect to subsets of another stream (e.g., $cam37$ with respect to $cam31$, Fig. 5).

On the other hand, by marginalizing over $k$, the distribution of the activities in the data set is recovered, i.e. $P(e) = \sum_k P(k, e)$. It can be noted that (cfr. Fig. 4) such distribution is not uniform: some activities are under-represented compared to other activities. This class imbalance problem entails two issues. First, any kind of processing performed to select subsets of the video streams for collating the most relevant data and information of interest should preserve the shape of such distribution, i.e. $\widetilde{P}(e) \approx P(e)$, where $\widetilde{P}(e)$ is the marginal distribution after processing. Second, non uniformity should be accounted for when defining quantitative evaluation measures [56]. Indeed, a suitable metric should reveal the true behavior of the method over minority and majority activities: the assessments of over-represented and under-represented activities should contribute equally to the assessment of the whole method. To cope with such a problem we jointly use two assessment metrics: the *standard accuracy* and the *macro average accuracy* [56].

Denote: $NP_e$ the number of positives, i.e., the number of times the activity $e$ occurs in the entire recorded scene, independently of the camera; $TP_e$ the number of *true positives* for activity $e$, i.e., the number of frames of the output video sequence that contain activity $e$. Given $NP_e$ and $TP_e$ for each activity, the following can be defined.

- *Standard Accuracy* $A = \frac{\sum_{e=1}^{E} TP_e}{\sum_{e=1}^{E} NP_e}$. Note that this is a global measure that does not take into account the accuracy achieved on a single activity. From now on we will refer to this metric simply as *accuracy*.
- *Macro Average Accuracy* $avg(A) = \frac{1}{E}\sum_{e=1}^{E} A_e = \frac{1}{E}\sum_{e=1}^{E} \frac{TP_e}{NP_e}$. This is the arithmetic average of the partial accuracy $A_e$ of each activity. It allows each partial accuracy to contribute equally to the method assessment. We will refer to this metric simply as *average accuracy*.

### D. Parameters and experiments setup

We used $1000$ frames to setup strategy parameters:

- *Bayesian*: the initial hyper-parameters $\nu_0^{(k)}, \Delta_0^{(k)}$;
- *Random*: the parameter $b_w$ of the probability distribution;
- *Deterministic*: the within-stream time parameter $\Delta_w$ that modulates camera switches;
- *Charnov*: the slope of the gain function $\delta$.

The remaining $7000$ frames have been used for testing giving-up strategies against the different visual representations previously introduced.

Note that a further constraint is to be taken into account for a fair performance assessment. In foraging terms, it is the number of times the forager chooses to explore a new patch; namely the number of camera switches.

While in general, the estimated parameters are those that maximize performance of a method, here parameters have been selected so to maximize the accuracy (or average accuracy) while keeping the number of camera switches below a given boundary condition. A measure of the accuracy of the system, should not be given as an absolute value, but the selection of subsets of the video streams should be performed to collate a "meaningful" summary in which the switching frequency is bounded. To thoroughly address this surmise, the accuracy behavior as function of the number of camera switches has been studied. The overall result can be summarised as: first, all representation schemes, apart from S, combined with the Bayesian giving-up strategy, achieve their best performance at a small number of camera switches; second, all giving-up strategies combined with the visual information method M achieve their best performance at a small number of camera switches (cfr., Fig. 14a and 14b of Supplementary Material). That being the experimental evidence, a reasonable upper bound can be determined either by taking into account the intrinsic limitations of the human visual system and/or, semantically, the characteristics of time activity distribution in the dataset. As to the first issue, consider that human subjects looking at videos explore the scene through saccadic eye-movements with maximal saccade duration of approximately $160$ ms and $340$ ms average post-saccadic fixational time (when cognitive processing takes place)[57]. Post-saccadic exploration can be even longer in case of pursuit (depending on task). Thus, a reasonable time to be granted for visual foraging is approximately one second (e.g, one/two saccades followed by pursuit, or two/tree control saccades with brief fixations). The length of the test scene is about $7000$ frames, $30$ fps frame rate, thus a conservative upper bound for the number of switches is about $240$. This is somehow consistent with empirical analysis of accuracy over switch number, where above $300$ camera switches strategies become comparable to the random strategy, in some cases worse. Under the circumstances, we slightly relax the upper bound to $280$.

As regards activity duration, note that the average length of each activity occurring in the scene across cameras is about $500$ frames. Ideally, a camera switch should take place after having observed a full activity. Thus, given a stream length of $7000$ frames, an upper bound estimate for the number of camera switches is about $14$. Since each boundary condition might determine a different set of parameters, distinct learning and testing phases have been performed for each boundary condition, that is #cam_switch $< 280$ and #cam_switch $< 14$.

### E. Results

Table II and Table III report quantitative assessment of results achieved by the different foraging strategies dependent on the available visual representations. Table II has been obtained by considering the upper bound #cam_switch $< 280$; Table III relates to the condition #cam_switch $< 14$.

Beyond the fact that the M/Bayesian scheme, at the core of the proposed model, overcomes other schemes both in terms of accuracy and average accuracy, some interesting results are worth a comment.

First, the proposed Bayesian giving-up strategy performs better than other strategies, in terms of both standard and average accuracy, independent of the visual representation adopted. At the same time, it is not affected by the chosen upper bound on the number of camera switches, whilst for other strategies the "semantic" upper bound (#cam_switch $< 14$) pairs with a slight decrease in performance. Both results

| Visual Information | Measure | Foraging strategy Random | Deterministic | Charnov | Bayesian |
|---|---|---|---|---|---|
| M | accuracy | 60.11 | 68.52 | 61.83 | **82.53** |
| | avg_accuracy | 45.77 | 52.09 | 48.20 | **71.12** |
| M+ | accuracy | 59.35 | 67.33 | 53.33 | 77.06 |
| | avg_accuracy | 43.95 | 58.32 | 45.55 | 57.56 |
| S | accuracy | 16.24 | 19.89 | 24.40 | 79.29 |
| | avg_accuracy | 14.33 | 19.45 | 14.98 | 70.79 |
| S+CD+FB | accuracy | 21.17 | 24.64 | 35.61 | 82.08 |
| | avg_accuracy | 19.36 | 24.56 | 21.95 | 68.47 |

TABLE II
ACCURACY AND AVERAGE ACCURACY ACHIEVED BY THE BASELINE AND PROPOSED FORAGING STRATEGIES COMBINED WITH SEVERAL VISUAL PROCESSING METHODS. RESULTS HAVE BEEN OBTAINED FOR #CAM_SWITCH $< 280$. BEST PERFORMANCE ARE REPORTED IN BOLD.

| Visual Information | Measure | Foraging strategy Random | Deterministic | Charnov | Bayesian |
|---|---|---|---|---|---|
| M | accuracy | 50.04 | 54.18 | 59.83 | **88.74** |
| | avg_accuracy | 33.08 | 29.60 | 34.62 | **76.99** |
| M+ | accuracy | 50.99 | 56.55 | 55.09 | 80.72 |
| | avg_accuracy | 34.06 | 35.49 | 29.50 | 72.98 |
| S | accuracy | 16.24 | 20.58 | 41.42 | 59.15 |
| | avg_accuracy | 14.33 | 14.98 | 22.85 | 50.54 |
| S+CD+FB | accuracy | 19.95 | 20.66 | 26.09 | 74.98 |
| | avg_accuracy | 16.23 | 14.91 | 19.48 | 51.05 |

TABLE III
ACCURACY AND AVERAGE ACCURACY FOR #CAM_SWITCH $< 14$.

confirm that for the task of monitoring multiple cameras, the stream switching strategy is a crucial issue, which might drastically affect the overall performance.

Second, the behavior of the monitoring system is not agnostic about the visual representation adopted. Results reported in both conditions, give quantitative evidence of the higher representation capability of M and M+ as opposed to S and S+CD+FB, which could be qualitatively appreciated by simple visual inspection of the $\mathcal{C}^{(k)}(t)$ behavior (graphs shown in Supplementary Materials). This holds independently of the strategy followed. Such effect is more clear by considering results obtained via the Random strategy. Recall that the latter selects a new stream after a within-stream time interval $\Delta_w$, which is randomly sampled. However, the camera content is not selected by chance and the quality of the visual complexity index $\mathcal{C}^{(k)}(t)$ plays a fundamental role for determining stream selection.

Third, the best performance achieved by M with respect to M+, shows that the ideal observer who behaves in an uncertain environment as an "omniscient forager" (switching is decided by surmising full knowledge of objects in streams [1]) is likely to perform less efficiently than a "prudent" one. This, together with the remarkable difference between results achieved via Charnov and Bayesian strategies, confirms the inadequacy of a pure Charnov strategy in complex experimental setting [4]. Similar considerations could be formulated on the other results reported as Supplementary Material.

At a glance, the overall behavior of the M/Bayesian scheme can be appreciated in Fig. 6, which is equivalent to the representation provided in Fig. 4, but computed on the output collated stream. The joint distribution $\widetilde{P}(k,e)$ and the marginal $\widetilde{P}(k)$ appear, as expected, to be a sparse summarisation of the initial dataset distributions $P(k,e)$ and $P(k)$ shown in Fig. 4. In particular, apart from *cam37*, the Bayesian strategy selects the most informative cameras, e.g., *cam16*, *cam31*, while discarding the less informative ones, e.g., *cam20*, *cam27*. Most important, the model provides an

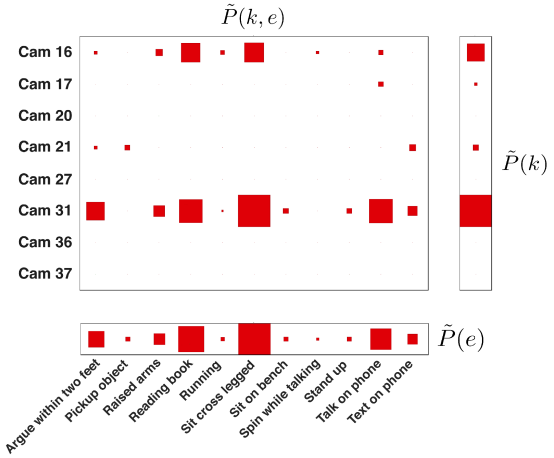


Fig. 6. Distribution of activities in terms of joint and marginal probability distributions $\tilde{P}(k,e)$, $\tilde{P}(k)$ and $\tilde{P}(e)$, respectively, obtained after processing (M representation and Bayesian foraging strategy with optimal #cam_switch $< 14$). To be compared with Fig. 4.

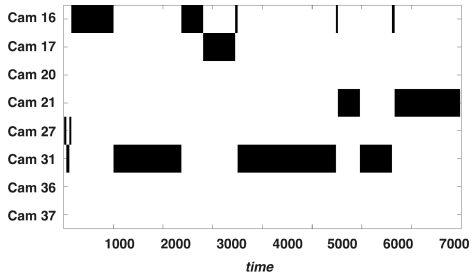


Fig. 7. Output timeline via the M/Bayesian scheme. It shows the sequence of the video stream subsets collated by the proposed model in terms of camera switches. At each time, the corresponding camera content is represented as a black rectangle and the output stream includes the content of one camera.

output marginal distribution of activities $\widetilde{P}(e)$ that is very close to the initial distribution $P(e)$. This result shows that by exploiting the proposed approach, most relevant activities have been captured.

Eventually, the concrete output of the M/Bayesian scheme's can be summarised in terms of the timeline visualized in Fig.7. This represents the final "storyboard", i.e., the video stream subsets collated by the proposed model in terms of camera switches. Here the output stream contains only one camera content at each time, and it can be considered as a "binarized" version of Fig. 5. From such representation, the example presented in Fig. 8 can be recovered. The sequence of most important frames can be thought of as a new composed video obtained by sequentially switching from one camera to another. As it can be observed (Fig. 8), each camera switch has been triggered by human activities, e.g. *walking* at t=1, *spin while talking* at t=2808, *raised arms* at at t=6661, etc.

## IX. FINAL REMARKS AND CONCLUSION

We have presented a unifying theoretical framework for selecting subsets of multiple video streams for collating the most relevant data and information of interest related to a given task. The framework formulates attentive monitoring as the behavior of a forager that, moment to moment, focuses


Fig. 8. A typical output of the M/Bayesian scheme, which recaps the foraging activity: the subset of frames extracted captures the most important data for the surveillance analysis task. Here, the output sequence is summarised through the camera switches performed by the optimal Bayesian observer. From the first at t=1 (top-left) to the last at t=6661 (bottom-right). In each camera frame the actual FoA is displayed as a white/blue circle.

his attention on the most informative stream/camera, detects interesting objects for the task at hand, switches from the current stream to a more informative one. Experimental results achieved on the UCR Videoweb Activities Dataset, have been presented to assess the performance of the proposed technique. To the best of our knowledge the model proposed here is novel for the multi-camera surveillance research field.

The approach could be either straightforwardly exploited for i) reducing the manual operator fatigue for multiple monitor situation or ii) as a preliminary step for intelligent surveillance relying on the analysis of actions, activities and behaviors. There are however some current limitations in the model that should be addressed for on-field application.

As to the first scenario, the actual visual foraging of a human operator should be taken into account for learning model parameters, which should entail two steps. First, mobile eye-tracking of operator's gaze behavior can be performed in the experimental setting of a typical control center with the human engaged in inspecting a number of camera monitors on the wall. There are few experiments of this sort in the psychological literature (e.g. [4]) but limited to simple target visual search. In the present work, eye-tracking data from human subjects have been used, but limited to the inference of parameters of distributions related to oculomotor biases; namely, the prior for sampling gaze shifts within the stream [40], [39]. Second, the components of the model should be implemented in order to allow full learning. For what regards the visual component (Fig.2), in a time-varying perspective, it can be conceived as a time slice of a Dynamic Bayesian Network; then, distribution parameters can be learned with a variety of methods available [45]. Foraging parameters of the executive component can be inferred using optimization methods that have been proposed for dealing with actual forager behaviors in a variety of patch/prey conditions [48].

For what concerns high-level intelligent analysis, current implementation of the model focuses on local object analysis and does not consider different levels of semantic information captured by cameras with different scales and angles. Attentive modeling of actions, activities and behaviors is a hot field in computer vision and results obtained up to now could be integrated within our framework with moderate effort. As it has been shown in the experimental analyses, the model offers a probabilistic framework in which it is easy to accommodate a variety of available state-of-the-art attention-based algorithms [26]. Further, note that the Bayesian strategy (Eq. 16) basically relies on the configurational complexity $\mathcal{C}^{(k)}(t)$, which, in turn, is based on spatial entropy. The latter is a mesoscopic quantity that summarises and can be derived from a variety of "atomic " visual measures (e.g, see [13]). However, from a strict engineering perspective much depends on the specific field of application that is to be addressed.

In the specific case of multi-stream summarisation, for instance, the method can be used as such, similarly to Kankanhalli *et al.* [31]. Alternatively, it is suitable to provide a principled base to approaches such as those performing correspondence-free multi-camera activity analysis [14].

An interesting issue is the applicability of the approach to the case of online multi-camera systems. This case compels to take into account the architectural complexities of the network. The latter can be factored in terms of distribution, mobility and degree of motion of the sensors [58]. As to the distribution issue, the pre-attentive control loop is suitable to be considered for a straightforward fully decentralized implementation, while the attentive loop could be designed at different levels of distribution. Interestingly enough, the ecological Bayesian handoff mechanism is suitable to embed resource-aware conditions, e.g., energy consumption, that are also considered in actual animal foraging. For what concerns the degree of motion of the sensors, clearly, the visual attention rationale that is behind our model calls for considering smart camera networks embedding PTZ cameras that are able to dynamically modify their FOV. In this case, for what concerns single camera activities, techniques developed in the active vision field are apt to be embedded in within-frame analysis either at the pre-attentive or the attentive stage [19]. However, at some point multi-camera activity analysis requires fusing information from multiple camera views. The data fusion problem has not been explicitly considered in this paper. Yet observational data may be combined, or fused, at a variety of levels [59], again depending on the architecture devised for a specific application.

In ongoing research we are considering multimodal data-fusion at the sensor level for audio/video integration in ambient intelligence. To such end the perceptual component can be straightforwardly extended to cope with other sources of information. Indeed, crossmodal integration can apparently arise before attentional selection is completed [60], which can be accounted for by exploiting the priority map for representing cross modal integration at this level. Addressing fusion at higher levels calls for software architecture abstractions to allow components to interact even if they rely on different spatial models. In this perspective, we are adapting a framework of space-based communication, to serve as an architectural

support for foraging in augmented ecologies [61].

## APPENDIX A
## PROOF OF PROPOSITION 7.1

For an optimal Bayesian observer the decision to leave the current stream is based on the posterior probability that a reward can be gained within the stream (complexity), given that no reward has been gained by time $t$; via Bayes' rule:

$$P(\mathcal{C}^{(k)}(t) \mid \neg R^{(k)}(t)) = \frac{P(\neg R^{(k)}(t) \mid \mathcal{C}^{(k)}(t))P(\mathcal{C}^{(k)}(t))}{P(\neg R^{(k)}(t))} \tag{18}$$

where $P(\neg R^{(k)}(t)) = 1 - P(R^{(k)}(t))$, $P(R^{(k)}(t))$ denoting the marginal likelihood of being rewarded. Using the detection function, Eq.15, the likelihood of not gaining reward is

$$P(\neg R^{(k)}(t) \mid \mathcal{C}^{(k)}(t)) = \exp(-\lambda t). \tag{19}$$

Since, by definition, reward can be actually gained only within the currently visited stream,

$$P(R^{(k)}(t)) = \sum_{\mathcal{C}(t)\in\{\mathcal{C}^{(k)}\}_{k=1}^{K}} P(\mathcal{C}(t))P(R^{(k)}(t) \mid \mathcal{C}(t)) = P(\mathcal{C}^{(k)}(t))P(R^{(k)}(t) \mid \mathcal{C}^{(k)}(t)). \tag{20}$$

Taking into account that $P(\neg R^{(k)}(t)) = 1 - P(R^{(k)}(t))$, the definition of the detection function, Eq. 15 and Eq. 20

$$P(\neg R^{(k)}(t)) = 1 - P(\mathcal{C}^{(k)}(t))(1 - \exp(-\lambda t)). \tag{21}$$

By the total law of probability, $1 - P(\mathcal{C}^{(k)}(t)) = \sum_{i\neq k} P(\mathcal{C}^{(i)}(t))$, thus previous equation can be written as

$$P(\neg R^{(k)}(t)) = \sum_{i\neq k} P(\mathcal{C}^{(i)}(t)) - P(\mathcal{C}^{(k)}(t)) \exp(-\lambda t). \tag{22}$$

Plugging into the posterior (Eq. 18) and rearranging

$$P(\mathcal{C}^{(k)}(t) \mid \neg R^{(k)}(t)) = \frac{P(\mathcal{C}^{(k)}(t))}{\exp(\lambda t)\sum_{i\neq k} P(\mathcal{C}^{(i)}(t)) - P(\mathcal{C}^{(k)}(t))}. \tag{23}$$

Optimal behavior consist in switching when the posterior is equal for all streams, thus

$$\frac{1}{K} = \frac{P(\mathcal{C}^{(k)}(t))}{\exp(\lambda t)\sum_{i\neq k} P(\mathcal{C}^{(i)}(t)) - P(\mathcal{C}^{(k)}(t))} \tag{24}$$

which gives the condition:

$$KP(\mathcal{C}^{(k)}(t)) = \exp(\lambda t)\sum_{i\neq k} P(\mathcal{C}^{(i)}(t)) + P(\mathcal{C}^{(k)}(t)). \tag{25}$$

Rearranging terms, using the prior probability $P(\mathcal{C}^{(k)}(t))$, Eq. 9, and inserting in Eq. 25, then the optimal condition for stream leaving, boils down to

$$\mathcal{C}^{(k)}(t) \exp(-\lambda t) = \frac{1}{K-1}\sum_{i\neq k} \mathcal{C}^{(i)}(t), \tag{26}$$

which proofs Eq. 16.

## ACKNOWLEDGMENTS

The authors are grateful to the Referees and the Associate Editor, for enlightening remarks that have greatly improved the quality of an earlier version of this paper, and to Prof. Raimondo Schettini for having inspired the present work.

## REFERENCES

[1] D. W. Stephens, *Foraging theory*. Princeton University Press, 1986.
[2] T. Xiang and S. Gong, "Beyond tracking: Modelling activity and understanding behaviour," *Int. J. Comput. Vis.*, vol. 67, no. 1, pp. 21–51, 2006.
[3] T. T. Hills, "Animal foraging and the evolution of goal-directed cognition," *Cognitive Science*, vol. 30, no. 1, pp. 3–41, 2006.
[4] J. M. Wolfe, "When is it time to move to the next raspberry bush? foraging rules in human visual search," *J. Vis.*, vol. 13, no. 3, p. 10, 2013.
[5] E. L. Charnov, "Optimal foraging, the marginal value theorem," *Theoretical population biology*, vol. 9, no. 2, pp. 129–136, 1976.
[6] A. Schütz, D. Braun, and K. Gegenfurtner, "Eye movements and perception: A selective review," *Journal of Vision*, vol. 11, no. 5, 2011.
[7] J. M. Fuster, "Upper processing stages of the perception–action cycle," *Trends in cognitive sciences*, vol. 8, no. 4, pp. 143–145, 2004.
[8] ——, "Cortex and memory: emergence of a new paradigm," *J. of Cognitive Neuroscience*, vol. 21, no. 11, pp. 2047–2072, 2009.
[9] S. Haykin and J. M. Fuster, "On cognitive dynamic systems: Cognitive neuroscience and engineering learning from each other," *Proc. IEEE*, vol. 102, no. 4, pp. 608–628, 2014.
[10] S. Chiappino, P. Morerio, L. Marcenaro, and C. S. Regazzoni, "Bio-inspired relevant interaction modelling in cognitive crowd management," *J. of Ambient Intell. and Humanized Computing*, pp. 1–22, 2014.
[11] S. Chiappino, L. Marcenaro, P. Morerio, and C. Regazzoni, "Event based switched dynamic bayesian networks for autonomous cognitive crowd monitoring," in *Wide Area Surveillance*. Springer, 2014, pp. 93–122.
[12] A. Dore, A. F. Cattoni, and C. S. Regazzoni, "Interaction modeling and prediction in smart spaces: a bio-inspired approach based on autobiographical memory," *IEEE Trans. Syst., Man, Cybern. A*, vol. 40, no. 6, pp. 1191–1205, 2010.
[13] S. Chiappino, L. Marcenaro, and C. Regazzoni, "Selective attention automatic focus for cognitive crowd monitoring," in *10th Int. Conf. Advanced Video Signal Based Surveillance*, Aug 2013, pp. 13–18.
[14] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 3 – 19, 2013.
[15] C. De Leo and B. S. Manjunath, "Multicamera video summarization and anomaly detection from activity motifs," *ACM Trans. on Sensor Networks*, vol. 10, no. 2, p. 27, 2014.
[16] S.-H. Ou, C.-H. Lee, V. Somayazulu, Y.-K. Chen, and S.-Y. Chien, "On-line multi-view video summarization for wireless video sensor network," *IEEE J. Select. Topics Signal Processing.*, vol. 9, no. 1, pp. 165–179, Feb 2015.
[17] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 66–75, 2012.
[18] R. Tron and R. Vidal, "Distributed computer vision algorithms," *IEEE Signal Processing Mag.*, vol. 28, no. 3, pp. 32–45, 2011.
[19] C. Micheloni, B. Rinner, and G. Foresti, "Video analysis in pan-tilt-zoom camera networks," *IEEE Signal Processing Mag.*, vol. 27, no. 5, pp. 78–90, Sept 2010.
[20] A. Kamal, C. Ding, A. Morye, J. Farrell, and A. Roy-Chowdhury, "An overview of distributed tracking and control in camera networks," in *Wide Area Surveillance*, ser. Augmented Vision and Reality, V. K. Asari, Ed. Springer Berlin Heidelberg, 2014, vol. 6, pp. 207–234.
[21] F. Qureshi and D. Terzopoulos, "Smart camera networks in virtual reality," *Proc. IEEE*, vol. 96, no. 10, pp. 1640–1656, 2008.
[22] Y. Li and B. Bhanu, "Utility-based camera assignment in a video network: A game theoretic framework," *IEEE Sensors J.*, vol. 11, no. 3, pp. 676–687, 2011.
[23] L. Esterle, P. R. Lewis, X. Yao, and B. Rinner, "Socio-economic vision graph generation and handover in distributed smart camera networks," *ACM Trans. on Sensor Networks*, vol. 10, no. 2, p. 20, 2014.
[24] D. Ballard, "Animate vision," *Art. Intell.*, vol. 48, no. 1, pp. 57–86, 1991.
[25] Y. Aloimonos, *Active perception*. Psychology Press, 2013.
[26] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 1, pp. 185–207, 2013.

[27] B. Dieber, C. Micheloni, and B. Rinner, "Resource-aware coverage and task assignment in visual sensor networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 10, pp. 1424–1437, Oct 2011.

[28] E. Sommerlade and I. Reid, "Probabilistic surveillance with multiple active cameras," in *Proc. IEEE ICRA*. IEEE, 2010, pp. 440–445.

[29] C. Ding, B. Song, A. Morye, J. Farrell, and A. Roy-Chowdhury, "Collaborative sensing in a distributed ptz camera network," *IEEE Trans. Image Processing*, vol. 21, no. 7, pp. 3282–3295, July 2012.

[30] A. Morye, C. Ding, A. Roy-Chowdhury, and J. Farrell, "Distributed constrained optimization for bayesian opportunistic visual sensing," *IEEE Trans. Contr. Syst. Technol.*, vol. 22, no. 6, pp. 2302–2318, Nov 2014.

[31] M. S. Kankanhalli, J. Wang, and R. Jain, "Experiential sampling on multiple data streams," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 947–955, 2006.

[32] N. Martinel, C. Micheloni, and G. L. Foresti, "Saliency weighted features for person re-identification," in *Proc. ECCV*, no. i, 2014, pp. 1–17.

[33] N. Ejaz, I. Mehmood, and S. W. Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Processing: Image Communication*, vol. 28, no. 1, pp. 34–44, 2013.

[34] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. IEEE CVPR*, June 2012, pp. 1346–1353.

[35] B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," in *Proc. IEEE CVPR*. IEEE, 2014, pp. 2513–2520.

[36] P. Napoletano and F. Tisato, "An attentive multi-camera system," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2014, pp. 90 240O–90 240O.

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

[38] J. K. Waage, "Foraging for patchily-distributed hosts by the parasitoid, nemeritis canescens," *The J. of Animal Ecology*, pp. 353–371, 1979.

[39] G. Boccignone and M. Ferraro, "Ecological sampling of gaze shifts," *IEEE Trans. Cybernetics*, vol. 44, no. 2, pp. 266–279, 2014.

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

[41] A. Clavelli, D. Karatzas, J. Lladós, M. Ferraro, and G. Boccignone, "Modelling task-dependent eye guidance to objects in pictures," *Cognitive Computation*, vol. 6, no. 3, pp. 558–584, 2014.

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

[43] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A bayesian inference theory of attention," *Vis. Res.*, vol. 50, no. 22, pp. 2233–2247, 2010.

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

[45] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. Cambridge, MA: MIT press, 2009.

[46] G. Boccignone, P. Campadelli, A. Ferrari, and G. Lipori, "Boosted tracking in video," *IEEE Signal Processing Lett.*, vol. 17, no. 2, pp. 129–132, 2010.

[47] A. Torralba, "Modeling global scene factors in attention," *JOSA A*, vol. 20, no. 7, pp. 1407–1418, 2003.

[48] J. M. McNamara, R. F. Green, and O. Olsson, "Bayes' theorem and its applications in animal behaviour," *Oikos*, vol. 112, no. 2, pp. 243–251, 2006.

[49] P. R. Killeen, G.-M. Palombo, L. R. Gottlob, and J. Beam, "Bayesian analysis of foraging by pigeons (*Columba livia*)." *J. Exp. Psychology: Animal Behavior Processes*, vol. 22, no. 4, p. 480, 1996.

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

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

[52] J. J. van Alphen and C. Bernstein, *Information Acquisition, Information Processing, and Patch Time Allocation in Insect Parasitoids*. Blackwell Publishing Ltd, 2008, pp. 172–192.

[53] G. Denina, B. Bhanu, H. T. Nguyen, C. Ding, A. Kamal, C. Ravishankar, A. Roy-Chowdhury, A. Ivers, and B. Varda, "Videoweb dataset for multi-camera activities and non-verbal communication," in *Distributed Video Sensor Networks*. Springer, 2011, pp. 335–347.

[54] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[55] P. Lundberg and M. Åström, "Functional response of optimally foraging herbivores," *J. Theoret. Biology*, vol. 144, no. 3, pp. 367–377, 1990.

[56] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*. John Wiley & Sons, 2013.

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

[58] R. Cucchiara and G. Gualdi, "Mobile video surveillance systems: An architectural overview," in *Mobile Multimedia Processing*, ser. Lecture Notes in Computer Science, X. Jiang, M. Ma, and C. Chen, Eds. Springer Berlin Heidelberg, 2010, vol. 5960, pp. 89–109.

[59] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proc. IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[60] J. Driver and C. Spence, "Crossmodal attention," *Current opinion in neurobiology*, vol. 8, no. 2, pp. 245–253, 1998.

[61] F. Tisato, C. Simone, D. Bernini, M. P. Locatelli, and D. Micucci, "Grounding ecologies on multiple spaces," *Pervasive and mobile computing*, vol. 8, no. 4, pp. 575–596, 2012.