Ferreira, J. F., Castelo-Branco, M., Dias, J., a hierarchical Bayesian framework for multimodal active perception, Adaptive Behavior, published online ahead of print, March 1st, 2012, doi: 10.1177/1059712311434662.

Adaptive Behavior

Adaptive Behavior
0(0) 1–18

DOI: 10.1177/1059712311434662
adb.sagepub.com
SAGE

A hierarchical Bayesian framework for multimodal active perception

João Filipe Ferreira[1], Miguel Castelo-Branco[2] and Jorge Dias[3]

Abstract

In this article, we present a hierarchical Bayesian framework for multimodal active perception, devised to be emergent, scalable and adaptive. This framework, while not strictly neuromimetic, finds its roots in the role of the dorsal perceptual pathway of the human brain. Its composing models build upon a common spatial configuration that is naturally fitting for the integration of readings from multiple sensors using a Bayesian approach devised in previous work. The framework presented in this article is shown to adequately model human-like active perception behaviours, namely by exhibiting the following desirable properties: high-level behaviour results from low-level interaction of simpler building blocks; seamless integration of additional inputs is allowed by the Bayesian Programming formalism; initial 'genetic imprint' of distribution parameters may be changed 'on the fly' through parameter manipulation, thus allowing for the implementation of goal-dependent behaviours (i.e. top-down influences).

Keywords

Multimodal active perception, hierarchical Bayes models, bioinspired robotics, human -robot interaction, emergence, scalability, adaptive behaviour

Introduction

Expectations regarding robotics have shifted from the development of automatic tools to aid or even replace humans in highly repetitive, simple, but physically demanding tasks, to the emergence of autonomous robots and vehicles, and finally to the development of service and social robots fronted by complex human–robot interfaces (HRI). Along this journey from *automaticity* to *autonomy*, one of the major obstacles has been the problem of the extraction of useful information about the external environment from sensory readings – in other words, *perception*.

Humans and other animals actively direct their sensors to unknown and also to interesting parts of the perceptual scene, so as to build up a mental map of the surrounding environment. One reason for saccades (i.e. rapid head-eye movements) is to move the senses so that redundant evidence can be accumulated about a scene, lowering the overall uncertainty of individual sensor measurements and using limited-scope sensorial resources more efficiently. In fact, although vision is arguably the main instigator of active perception in mammals, this process is undoubtedly multisensory. As a remarkable example, audition is able to drive gaze shifts towards targets outside the visual field, an ability that has made this sense paramount for the interaction between humans and their surroundings. *Active perception* has been an object of study in robotics for decades now, specially active vision, which was first introduced by Bajcsy (1985) and later explored by Aloimonos, Weiss, & Bandyopadhyay (1987). Many perceptual tasks tend to be simpler if the observer actively shifts attention by controlling its sensors (Aloimonos et al., 1987). Active perception is thus an intelligent data acquisition process driven by the measured, partially interpreted scene parameters and their errors from the scene. The active approach has the important advantage of making most ill-posed perception tasks tractable (Aloimonos et al., 1987).

Since humans are prevalently social beings, their attentional system is inherently socially driven; this becomes particularly important when considering

[1]Institute of Systems and Robotics, FCT-University of Coimbra, Coimbra, Portugal
[2]Biomedical Institute of Research on Light and Image, Faculty of Medicine, University of Coimbra, Coimbra, Portugal
[3]Institute of Systems and Robotics, FCT-University of Coimbra, Coimbra, Portugal

Corresponding author:
João Filipe Ferreira, Institute of Systems and Robotics, FCT-University of Coimbra, Coimbra, Portugal
Email: jfilipe@isr.uc.pt

human–machine interaction, where robots are expected to engage with humans while displaying attentional behaviours that resemble those of their interlocutors. Even when dealing with unknown environments with no social intent, humans use their own attentional system to its full. For example, a random exploratory strategy alone would not take into account potential primal dangers lurking in our surroundings. In fact, human evolution has genetically imprinted as prior knowledge that certain stimuli are a tell-tale of the proximity of predators, or are caused by competitors of our own species. Consequently, Kopp & Gärdenfors (2002) posit that the capacity of attention, and therefore active perception, is a *minimal criterion of intentionality* for robots.

One of the most popular computational models serving as a basis for robotic implementations of visual attention is the *saliency model* by Itti, Koch, & Niebur (1998). This model has roots at least as far back as Niebur, Itti, & Koch (1995) and its most recent developments are described in Carmi & Itti (2006). The gaze computation process takes, as an input, the saliency map, and returns, as an output, a point of fixation. On the other hand, regarding the *temporal* dimension of attention, a commonly used complementary model is the *Inhibition of Return* (IoR) mechanism (Niebur et al., 1995). The IoR, in simple terms, is the mechanism where the saccade generating system in the brain avoids fixation sites which have just been a focus of attention, therefore preventing deadlocks and infinite loops.

Both humans and robots alike operate in a world of sensory uncertainty. Perception has as of recently, consequently and perhaps unsurprisingly, been considered increasingly as a computational process of unconscious, probabilistic inference (Knill & Pouget, 2004). Recent advances both in statistics and artificial intelligence have spurred researchers to begin to apply the concepts of probability theory rigorously to problems in biological perception and action (Doya, Ishii, Pouget, & Rao, 2007; Knill & Pouget, 2004). At the same time, researchers in the robotics community have also begun to apply probabilistic approaches for the development of models for artificial perception.

The perceptual process is inherently highly intricate and complex. To deal with this challenge, it is believed that, during evolution of the animal brain, this process has been decomposed into simpler subtasks, carried out by modular sites intertwined in a complex fashion, forming a myriad of forward, lateral and feedback connections. Current trends in probabilistic modelling of perception have introduced approaches such as *Hierarchical Bayes Models*, which allow for this kind of modular decomposition of complex processes into simpler models through the appropriate use of intermediate variables and independence assumptions.

It would be difficult to assume that natural cognitive systems process their complex sensory inputs in a single layer of computation (Colas, Diard, & Bessière, 2010; Lee, 2011). Therefore perception can be decomposed into subprocesses that communicate intermediate results, which introduces the notion of *modularity*.

Ever since seminal work by Marr (1982) up until more recent accounts such as Ballard (1999) and many others on computational theories of perception, the link between the functional organization of perceptual sites in the brain and the underlying computational processes has led to the notion that modularity plays a major role in making these processes tractable. As a matter of fact, although the interconnections between these sites have increasingly been found to be much more intricate than Marr believed, the notion that the brain is organized in a modular fashion is undisputed.

Hierarchical Bayesian methods are standard and powerful tools for analysing models and drawing inferences, and have been extensively applied in statistics, machine learning and throughout the empirical sciences (Shiffrin, Lee, Wagenmakers, & Kim, 2008; Lee, 2011). Hierarchical Bayesian methods provide the adequate framework for implementing modularity in perception. Firstly, these methods allow model development to take place at multiple levels of abstraction. Secondly, they offer the possibility of understanding emergent behaviour as resulting from a mixture of qualitatively and quantitatively different sources. And, thirdly, this framework is able to unify disparate models. There does not seem to be a theoretical consensus on the definition of what makes a model 'hierarchical' – see Lee (2011) for a detailed discussion on this matter. Therefore, it is important to make clear what is meant when referring to hierarchical Bayes methods: in the case of the models presented in this paper, we claim they are hierarchical in the broader sense, given that we accept that any model with more than one level of dependency between variables as such, but with the very important restriction that these frameworks should also be modular, that is that any of their composing models may be reworked while keeping the remaining structure untouched.

Formally, hierarchical Bayes models can be constructed through the use of probabilistic versions of subroutine calls (i.e. a probability distribution used in one model is given by the answer to a question to another model), conditional switches and weighting (probabilistic mixture models), and also through the use of Bayesian model abstraction, as described by Colas et al. (2010).

The use of more than one sensor promotes a robustness increase on the observation and characterization of a physical phenomenon. In fact, using different types of sensors allows for the dilution of each sensor's individual weaknesses through the use of the strengths of the remainder. Fortunately, the probabilistic approach provides the appropriate set of tools to perform sensor fusion while taking into account the effects of multimodality in the lowering of the overall uncertainty underpinning perception.

Overall goals and related work

We will present a complex artificial active perception system that follows human-like bottom-up driven behaviours using vision, audition and vestibular sensing.

We mainly expect to contribute with a solution which:

- deals with perceptual uncertainty and ambiguity, offering some adaptive ingredients that form a reasonable bioinspired basis for a full-fledged robotic perception system;
- deals with multimodality by tackling sensor fusion geometry in a natural way, consistent with most of the inherent properties of sensation;
- allows for fast processing of perceptual inputs to build a spatial representation for active perception in a behaviourally relevant fashion, as required in applications in which complex human–robot interaction is required.

The focus of this article will be concentrated on this last aspect, for which the Bayesian Programming formalism (Bessière, Laugier, & Siegwart, 2008) was applied to develop a hierarchical modular probabilistic framework that allows the combination of active perception behaviours, namely:

- active exploration based on entropy developed in previously published work, using a Bayesian filter operating upon a log-spherical occupancy grid, which, while not strictly neuromimetic, finds its roots in the role of the dorsal perceptual pathway and superior colliculus of the human brain – refer to Ferreira, Pinho, & Dias (2008a) and Ferreira, Prado, Lobo, & Dias (2009) for more details;
- automatic orientation based on sensory saliency (Niebur et al., 1995), also operating upon the same log-spherical grid.

A real-time implementation of all the processes of the framework has been developed, capitalizing on the potential for parallel computing of most of its algorithms, as an extension of what was presented in Ferreira, Lobo, & Dias (2011).

An overview of the framework and its models will be summarized in this text, and results will be presented. In the process, we will demonstrate the following properties which are intrinsic to the framework: *emergence, scalability* and *adaptivity*. This will be the realization of a general paradigm for multimodal perception research, both for neuroscience and for robotics, where the Bayesian framework, not only constitutes the formalism for model construction, but also defines the basis by which robot and human perception is compared, as shown on Figure 1.

Recent work in active vision by Tsotsos & Shubina (2007) and Bohg et al. (2009), the former for target search and the latter for object grasping, contrary to our solution, use an explicit representation for objects to implement active perception. On the other hand, several solutions for target applications similar to ours avoid explicit object representation by resorting to a bottom-up saliency approach such as defined by Itti et al. (1998) – examples of these would be Shibata, Vijayakumar, Conradt, & Schaal (2001), Breazeal, Edsinger,

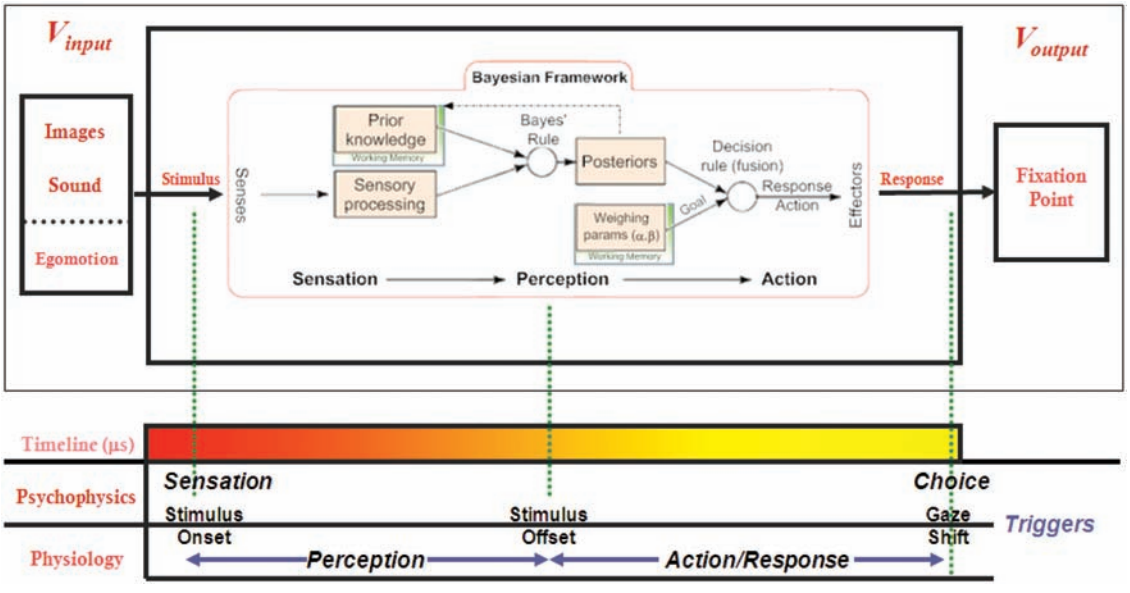

Figure 1. Experimental procedure schematic. The tentative model is interchangeable with a human observer and performances can thus be compared. Preliminary knowledge on the environment, context and current task is encoded in the prior and the decision parameters of the model, both of which are stored in working memory, as shown. Experimental techniques may include, but are not limited to, psychophysics, electrophysiology, fMRI, body tracking, etc.

Fitzpatrick, & Scassellati (2001) and Dankers, Barnes, & Zelinsky (2007). Finally, Dankers, Barnes, & Zelinsky (2005) use an approach similar to ours, with an egocentric three-dimensional (3D) occupancy grid for integrating range information using active stereo and a Bayesian approach, also detecting 3D mass flow. However, this solution suffers from the downside of using an Euclidean tesselation of space, which complicates sensor models for map updating and fixation computation due to the compulsory use of ray-tracing methods. These works, as most solutions in active perception, use a behavioural approach; an alternative is a probabilistic approach that attempts to reduce uncertainty on a part of the world state, modelled as belief (Croon, Sprinkhuizen-Kuyper, & Postma, 2009). Our work intends to combine both variants into a coherent, albeit more powerful, approach.

Active multisensory perception using spatial maps has, contrastingly, been the object of study for only a short time. Few other explicit models exist, although many artificial perception systems include some kind of simple attention module that drives gaze towards salient auditory features. As an example of a fully-fledged multisensory attention model, Koene, Morén, Trifa, & Cheng (2007) present a general architecture for the perceptual system of a humanoid robot featuring multisensory (audiovisual) integration, bottom-up salience detection, top-down attentional feature gating and reflexive gaze shifting, which is of particular relevance to our work. The complete system focuses on the multisensory integration and desired gaze shift computation performed in the 'Superior Colliculus (SC)' module (Koene et al., 2007). This allows the robot to orient its head and eyes so that it can focus its attention on audio and/or visual stimuli. The system includes mechanisms for bottom-up stimulus salience based gaze/attention shifts (where salience is a function of feature contrast) as well as top-down guided search for stimuli that match certain object properties. In order to facilitate interaction with dynamic environments the complete perceptual-motor system functions in real-time (Koene et al., 2007).

Our approach implements active visuoauditory perception, adding to it vestibular sensing/proprioception so as to allow for sensor fusion given a rotational egomotion. However our solution differs from purely saliency-based approaches in that it also implements an active exploration behaviour based on the entropy of the occupancy grid, so as to promote gaze shifts to regions of high uncertainty.

A Bayesian hierarchy as a model of active visuoauditory perception

Formalisms and notation for model construction

The models presented in this article will be described using the Bayesian Program formalism, as first defined by Lebeltel (1999) and later consolidated by Bessière et al. (2008). This formalism was created to supersede, restate and compare numerous classical probabilistic models such as Bayesian Networks, Dynamic Bayesian Networks, Bayesian Filters, Hidden Markov Models, Kalman Filters, Particle Filters, Mixture Models or Maximum Entropy Models.

Some important related notation issues will be presented next for easier reading of the text that follows.

- Random variables are represented in uppercase, such as C, and their instantiations are represented in lowercase, as in c. These instantiations are fully stated by proceeding as in the example that follows: $[C = c]$.
- π represents preliminary knowledge over the context consubstantiated by hidden and latent variables (i.e. intentionally or unintentionally unaccounted for factors). Preliminary knowledge notation $\pi_i, i \in [1, m]$ is generally used to distinguish between the context of m different models.
- To simplify reading and introduce homogeneity into the notation, single probability values, probability distributions and families of probability distributions will all generically be formally denoted as conditional probabilities, $P(\bullet \mid \bullet\, \pi)$. They are distinguished from one another by the context of their arguments. For simplicity, this notation can be reduced to $P(\bullet \mid \bullet)$ by making the influence of hidden and latent variables implicit.
- In exceptional cases, there are only dependences on hidden or latent variables, in which case the notation reduces to $P(\bullet \mid \pi)$, or more simply to $P(\bullet)$.
- Using this notation, $P(A|B)$ is a family of distributions, one for each possible value of B; $P(A|b)$ is one such distribution; $P(a|b)$ is a single probability value.
- Single probabilities can exceptionally and abusively also be denoted as P_{idx}, where idx may be any descriptive text, for easier reading.

Models

A spatial representation framework for multimodal perception of 3D structure and motion, the Bayesian Volumetric Map (BVM), was presented in Ferreira, et al. (2008). This framework is characterized by an egocentric, log-spherical spatial configuration to which the Bayesian Occupancy Filter (BOF), as formalized by Tay, Mekhnacha, Chen, Yguel, & Laugier (2008), has been adapted. It effectively provides a computational means of storing and updating a perceptual spatial map in a short-term working memory data-structure, representing both 3D structure and motion, without the need for any object segmentation process (see Figure 2), finding its roots in the role of the superior colliculus and the dorsal perceptual pathway of the human brain.

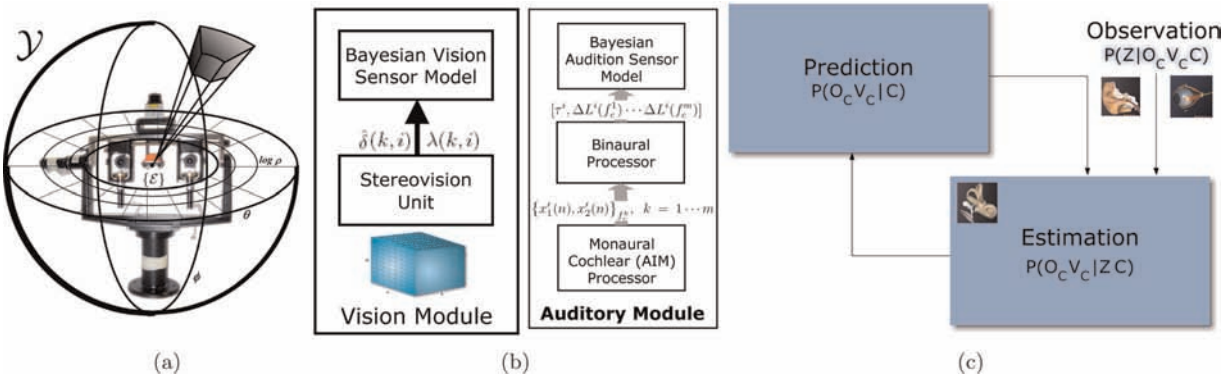

Figure 2. Multimodal perception framework details. (a) The Bayesian Volumetric Map (BVM) referred to the egocentric coordinate frame of the robotic active perception system; (b) BVM sensor models; (c) BVM Bayesian Occupancy Filter.

In the BVM framework, cells of a partitioning grid on the BVM log-spherical space $\mathcal{Y}$ associated with the egocentric coordinate system $\{\mathcal{E}\}$ are indexed through $C \in \mathcal{Y}$, representing the subset of positions in $\mathcal{Y}$ corresponding to the 'far corners' $(\log_b \rho_{\max}, \theta_{\max}, \phi_{\max})$ of each cell C, O_C is a binary variable representing the state of occupancy of cell C (as in the commonly used occupancy grids – see Elfes (1989)), and V_C is a finite vector of random variables that represent the state of all local motion possibilities used by the prediction step of the Bayesian filter associated to the BVM for cell C, assuming a constant velocity hypothesis, as depicted on Figure 2. Sensor measurements (i.e. the result of visual and auditory processing) are denoted by Z – observations $P(Z|O_C V_C C)$ are given by the Bayesian sensor models of Figure 2, which yield results already integrated within the log-spherical configuration.

The complete set of variables that set up the framework and its extensions, which will be described in the final part of this section, is summarized in the following list (references to temporal properties have been removed for easier reading):

- C: cell index on the BVM occupancy grid given by the 3D coordinates of its 'far corner';
- Z: generic designation for either visual or auditory sensor measurements;
- O_C: binary value signalling the fact that a cell C is either empty or occupied by an object;
- V_C: discrete variable indicating instantaneous local motion vector for objects occupying cell C;
- G: fixation point for next gaze-shift in log-spherical coordinates;
- U_C: entropy gradient-based variable ranging from 0 to 1, signalling the potential interest (i.e. 0 and 1 meaning minimally and maximally interesting, respectively) of cell C as future focus of attention given the uncertainty on its current state given by (O_C, V_C), thus promoting an active exploration behaviour;
- S_C^i: binary value describing the ith of N sensory saliency of cell C;
- $Q_C^i = P([S_C^i = 1]|Z^i C)$: probability of a perceptually salient object occupying cell C;
- R_C: inhibition level for cell C as a possible future focus of attention modelling the IoR behaviour, ranging from no inhibition (0) to full inhibition (1).

By restricting egomotion to rotations around the egocentric axes, vestibular sensing (see the following subsection), together with the encoders of the motors of the robotic head (i.e. proprioception), will yield measurements of angular velocity and position which can then be easily used to manipulate the BVM, which is, by definition, in spherical coordinates (Ferreira, Bessière, et al., 2008). In this case, the most effective solution for integration is to perform the equivalent index shift. This process is described by redefining C: $C \in \mathcal{Y}$ *indexes a cell in the BVM by its far corner, defined as* $C = (\log_b \rho_{\max}, \theta_{\max} + \theta_{\text{inertial}}, \phi_{\max} + \phi_{\text{inertial}}) \in \mathcal{Y}$.

For Bayesian stereovision sensing, we have decided to use a data structure loosely based on the neuronal population activity patterns found in the perceptual brain to represent uncertainty in the form of probability distributions (Pouget, Dayan, & Zemel, 2000). Thus, a spatially organized two-dimensional (2D) grid may have each cell associated to a 'population code' extending to additional dimensions, yielding a set of probability values encoding a N-dimensional probability distribution function. This information, conveniently expressed in cyclopean coordinates[1] (thus using the egocentric frame of reference $\{\mathcal{E}\}$), is consequently used as soft evidence[2] by a Bayesian sensor model previously presented in Ferreira, Bessière, et al. (2008) and Ferreira, Pinho, & Dias (2008b) – see Figure 2(b), on the left.

The Bayesian binaural system, which was fully described in Pinho, Ferreira, Bessière, & Dias (2008) and Ferreira, Pinho, & Dias (2009), is composed of three distinct and consecutive processors (Figure 2(b),

on the right): the monaural cochlear unit, which processes the pair of monaural signals $\{x_1, x_2\}$ coming from the binaural audio transducer system by simulating the human cochlea, so as to achieve a tonotopic representation (i.e. a frequency band decomposition) of the left and right audio streams; the binaural unit, which correlates these signals and consequently estimates the binaural cues and segments each sound source; and, finally, the Bayesian 3D sound source localization unit, which applies a Bayesian sensor model so as to perform localization of sound sources in 3D space, again using the egocentric frame of reference $\{\mathcal{E}\}$. This sensor model uses an intermediate variable, S_C, which signals the presence or absence of a sound source within cell C, and relates it to O_C through $P(S_C|O_C)$. Therefore, the same model can be also used to infer $P([S_C = 1]|ZC)$, the probability of a *sound-producing* object occupying cell C.

The BVM is extendible in such a way that other properties, characterized by additional random variables and corresponding probabilities might be represented. To this end, other than the already implemented occupancy and local motion properties O_C and V_C, additional properties were implemented by augmenting the hierarchy of operators through Bayesian subprogramming (Bessière et al., 2008; Lebeltel, 1999).

One such property that we propose to model uses the knowledge from the BVM to determine gaze shift fixation sites. More precisely, it elicits gaze shifts towards locations of high entropy/uncertainty based on the rationale conveyed by an additional variable that quantifies the uncertainty-based interest of a cell on the BVM, thus promoting entropy-based active exploration.

Therefore, we introduce a new random variable U_C, which takes the algorithm presented in Ferreira et al. (2008a) and expresses it in a compact mathematical form

$$U_C = \begin{cases} (1 - P([O_C = 1]|C)) \frac{\|\vec{\nabla} H(C)\|}{\max \|\vec{\nabla} H(C)\|} & C \in \mathcal{F} \\ 0 & C \notin \mathcal{F} \end{cases} \quad (1)$$

where $\mathcal{F} \subset \mathcal{Y}$ represents the set of *frontier cells* (cells belonging to a particular line-of-sight (θ_C, ϕ_C) in the BVM, just preceding the first occupied cell in that direction), and

$$\begin{aligned} H(C) &\equiv H(V_C, O_C) \\ &= -\sum_{O_C, V_C} P(O_C V_C|C) \log P(O_C V_C|C) \end{aligned}$$

and

$$\begin{aligned} &\| \vec{\nabla} H(C) \| = \\ &\| [H(C) - H(C_{\rho-}), H(C) - H(C_{\theta-}), H(C) - H(C_{\phi-})]^T \| \end{aligned}$$

represent the generic expressions of the *joint entropy* and *joint entropy gradient* of a cell C, respectively, as defined in Ferreira et al. (2008a). This implies that $U_C \in [0, 1]$, being close to 1 when uncertainty is high and C is a frontier cell, and $U_C \to 0$ when uncertainty is low or C is not a frontier cell. In the current implementation, we use the maximum value for this variable for the decision on the gaze shift, as described in Ferreira et al. (2008a).

To achieve our goal of designing Bayesian models for visuoauditory-driven saccade generation following human active perception behaviours, a hierarchical framework, inspired on what was proposed by Colas, Flacher, Tanner, Bessière, & Girard (2009), has been developed and is presented in the following text.

We will specify three decision models: π_A, that implements entropy-based active exploration based on the BVM and the heuristics represented by equation (1), π_B, that uses entropy and saliency together for active perception, and finally π_C which adds a simple IoR mechanism based on the fixation point of the previous time-step. In other words, each model incorporates its predecessor through Bayesian fusion, therefore constituting a model hierarchy – see Figure 3.

The hierarchy is extensible in such a way that other properties characterized by additional random variables and corresponding probabilities might be represented, other than the already implemented occupancy and local motion properties of the BVM, by augmenting the hierarchy of operators through Bayesian subprogramming (Bessière et al., 2008; Lebeltel, 1999). This ensures that the framework is *scalable*. On the other hand, the combination of these strategies to produce a coherent behaviour ensures that the framework is *emergent*.

Furthermore, each model will infer a probability distribution on the next point of fixation for the next desired gaze shift represented by a random variable $G^t \in \mathcal{Y}$ at each time $t \in [1, t_{\max}] : P(G^t|V^{1\to t} O^{1\to t} \pi_k)$, where $V^{1\to t} = \wedge_{t\in[1,t_{\max}]} \wedge_C V_C^t$ and $O^{1\to t} = \wedge_{t\in[1,t_{\max}]} \wedge_C O_C^t$ represent the conjunction of BVM local motion and occupancy estimate states for all cells $C \in \mathcal{Y}$, from system startup up until the current time-instant t.

The first model we propose uses the knowledge from the BVM layer to determine gaze shift fixation points. More precisely, it tends to look towards locations of high entropy/uncertainty. Its likelihood is based on the rationale conveyed by the additional variable U_C, defined earlier.

The Bayesian Program for this model is presented on Figure 4. The dependency of the uncertainty measure variable U_C^t – equation (1) – on the BVM states $(V^{1\to t}, O^{1\to t})$ are implicitly stated by definition; thus, with this model, the distribution on the point of fixation of the next desired gaze shift can be computed using the following expression

$$\begin{aligned} P(G^t|V^{1\to t} O^{1\to t} \pi_A) &= P(G^t|U^t \pi_A) \\ &\propto \prod_C P(U_C^t|G^t \pi_A) \end{aligned} \quad (2)$$

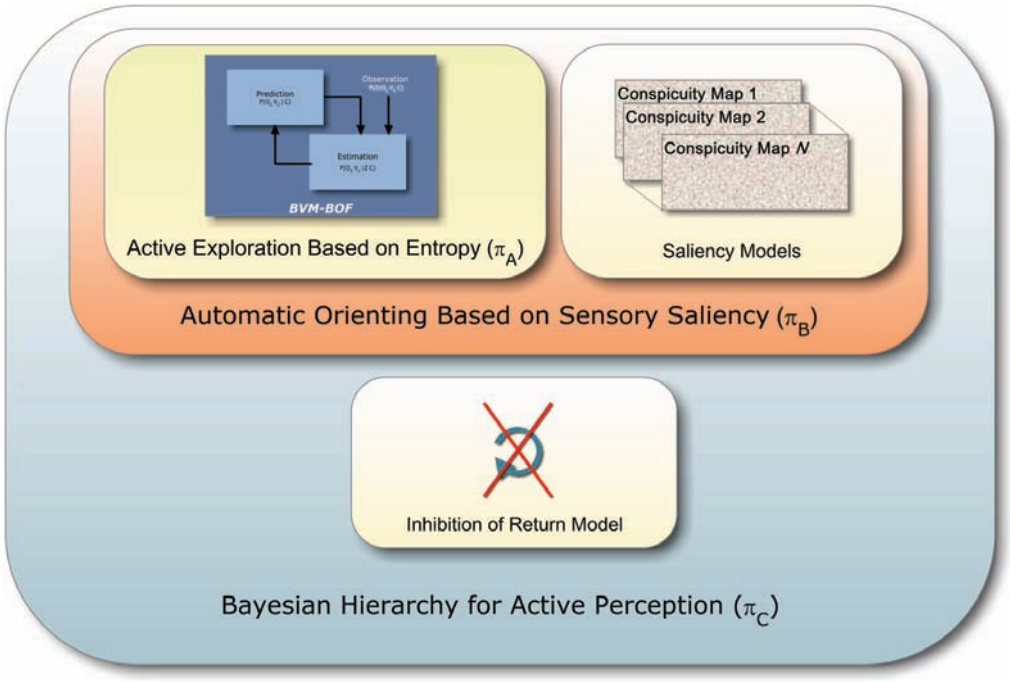

Figure 3. Conceptual diagram for active perception model hierarchy.

Program

Description

Specification

Relevant variables:

O_C^t: binary value describing the occupancy of cell C at time t, $[O_C = 1]$ if cell C is occupied by an object, $[O_C = 0]$ otherwise (related variable: $O^t = \bigwedge_C O_C^t$);

V_C^t: velocity of cell C at time t, discretized into $n+1$ possible cases $\in \mathcal{V} \equiv \{v_0, \cdots, v_n\}$ (related variable: $V^t = \bigwedge_C V_C^t$);

$G^t \equiv (\log_b \rho_{\max}, \theta_{\max}, \phi_{\max}) \in \mathcal{Y}$: fixation point for next gaze-shift, computed at time t, with related variable $G = \bigwedge_{t \in [1, t_{\max}]} G^t$;

$U_C^t \equiv f(V^{1 \to t}, O^{1 \to t}) \in [0, 1]$: joint entropy gradient-based variable at time t, with related variables $U^t = \bigwedge_C U_C^t$ and $U = \bigwedge_{t \in [1, t_{\max}]} U^t$

(see equation (1): close to 1 when uncertainty is high and C is a frontier cell, and $U_C \to 0$ when uncertainty is low or C is not a frontier cell).

Decomposition:

$$P(G\,U|\pi_A) = \prod_{t=1}^{t_{\max}} \left[P(G^t|\pi_A) \prod_C P(U_C^t|G^t\,\pi_A) \right]$$

Parametric forms:

$P(G^t|\pi_A)$: uniform prior;

$P(U_C^t|G^t\,\pi_A)$: is a beta distribution $\mathcal{B}(\alpha_U, \beta_U)$ for $[G^t = C]$ that expresses that, for a given point of fixation proposal for the next gaze shift, U_C^t is more likely near 1, or a uniform distribution on U_C^t for $[G^t \neq C]$.

Identification:

Empirical values for free parameters α_U and β_U.

Questions:

$$P(G^t|V^{1 \to t}\,O^{1 \to t}\,\pi_A) = P(G^t|U^t\,\pi_A)$$

Figure 4. Bayesian Program for entropy-based active exploration model π_A.

Figure 5. Bayesian Program for automatic orienting based on sensory saliency model π_B.

The second model is based on sensor models that relate sensor measurements $Z_j^{i,t}$ with $i = 1..N$ independent sensory properties of saliency ($j = 1..M_i$ total independent measurements for each saliency property), represented by the set of binary random variables $S_C^{i,t}$ (equalling 0 when the cell is non-salient and 1 when salient) corresponding to each cell C. In other words, these sensor models are generically notated as $P(Z^t|S_C^{i,t} V_C^t O_C^t \pi_C)$, indiscriminately of what the specific sensory saliency property $S^{i,t} = \wedge_C S_C^{i,t}$ might represent.

The Bayesian Program for model π_B is presented on Figure 5. With this model, the distribution on the point of fixation of the next desired gaze shift can be computed using the following expression

$$P(G^t|V^{1\to t} O^{1\to t} S^t\,\pi_B) \propto P(G^t|V^{1\to t} O^{1\to t} \pi_A) \prod_C \left[\prod_{i=1}^N P(Q_C^{i,t}|G^t\,\pi_B) \right] \quad (3)$$

In short, this model is the product between the prior on gaze shifts due to entropy-based active exploration and each distribution on the sensory-salient cells. This expression shows that the model is attracted towards both salient cells (without necessarily looking at one in particular, as the balance between the distributions on salient cells can lead to a peak in some weighted sum of their locations) *and* locations of high uncertainty when sensory saliency is not preponderant enough (i.e. this process is called *weighting*, as opposed to *switching*, in which these behaviours would be mutually exclusive – see Colas et al. (2010) and Ferreira & Castelo-Branco (2007)).

The Bayesian Program for the third and final model π_C, which defines the full active perception hierarchy by adding an implementation the IoR mechanism, is presented on Figure 6. With this model, the distribution on the point of fixation of the next desired gaze shift can be computed using the following expression

Figure 6. Bayesian Program for full active perception model π_C.

$$P(G^t|V^{1\to t}\,O^{1\to t}\,S^t\,G^{t-1}\,\pi_C) \propto P(G^t|V^{1\to t}\,O^{1\to t}\,S^t\,\pi_B) \prod_C \left[P(R_C^t|G^t\,\pi_C)\right] \quad (4)$$

In conclusion, the full hierarchy, represented graphically in Figure 7, is defined as the product between the prior on gaze shifts due to entropy-based active exploration and each distribution on the sensory-salient cells, while avoiding the fixation site computed on the previous time step through the IoR process, implemented by the last factor in the product. The parameters used for each distribution in this product, which define the relative importance of each level of the hierarchy and of each sensory saliency property, may be introduced directly by the programmer (like a genetic imprint) or manipulated 'on the fly', which in turn allows for goal-dependent behaviour implementation (i.e. top-down influences), therefore ensuring that the framework is *adaptive*.

Implementation of an artificial Bayesian active perception system

BVM–IMPEP framework implementation

The BVM–IMPEP framework, of which an implementation diagram is presented in Figure 8, was realized as follows: AQ2

- *Vision sensor system.* With the OpenCV toolbox and the implementation by Gallup (2009) of a basic binocular stereo algorithm on GPU using CUDA. The algorithm reportedly runs at 40 Hz on 640×480 images while detecting 50 different levels of

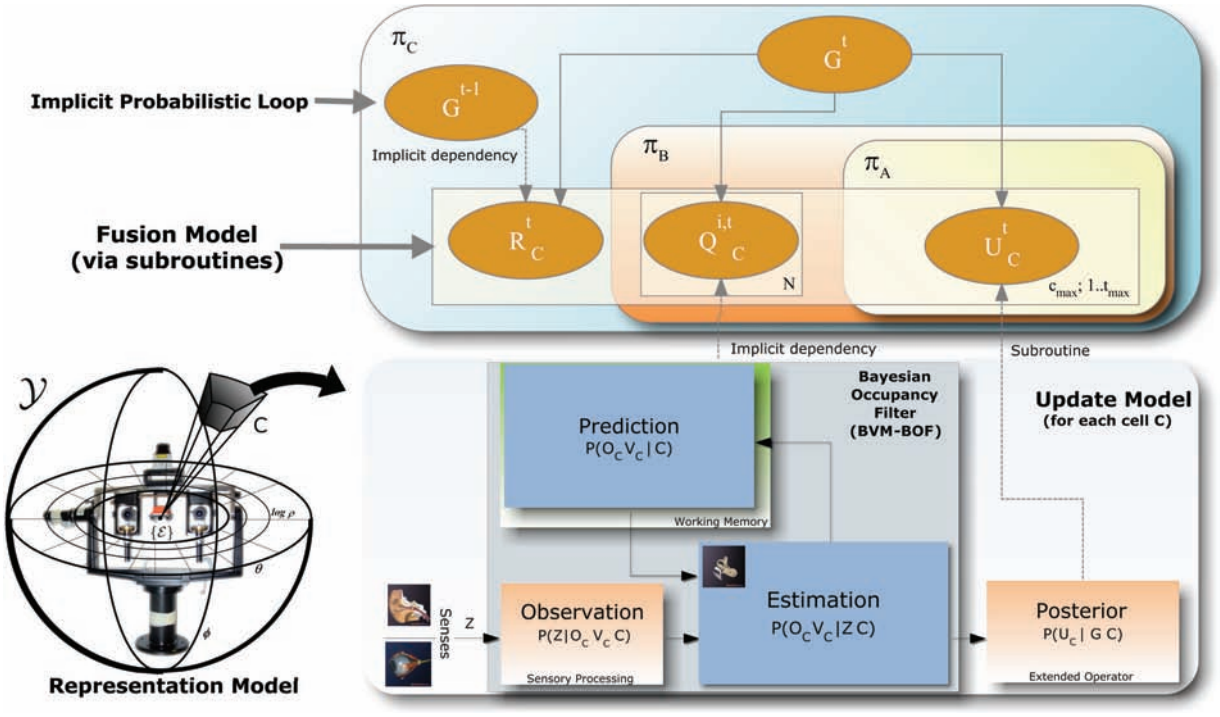

Figure 7. Graphical representation of the hierarchical framework for active perception. Bottom half: Update and Representation Models for the BVM–BOF framework, extended by the entropy gradient-based operator. Upper half: Bayesian network summarizing the models presented in this text, using the plates notation (an intuitive method of representing variables that repeat in a graphical model, so that the respective distributions appear in the joint distribution as an indexed product of the sequence of variables – for more information refer to Buntine (1994)). As can be seen, emergent behaviour results from a probabilistic fusion model implemented through a sequence of Bayesian Programming subroutines and an implicit loop that ensures the dynamic behaviour of the framework (Colas et al., 2010).

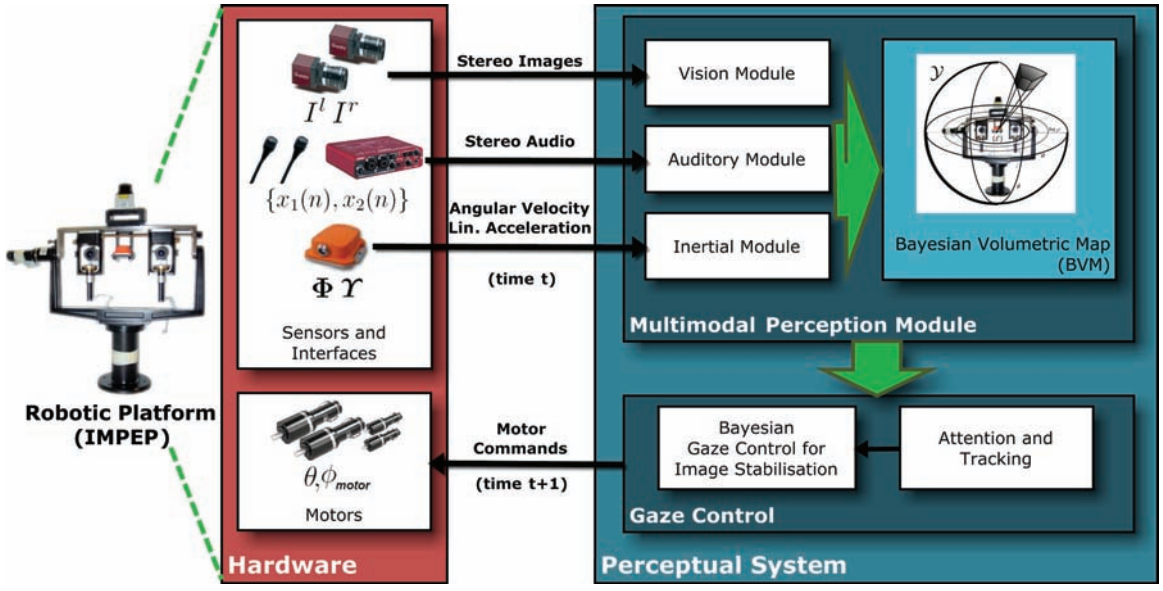

Figure 8. Implementation diagram for the BVM-IMPEP multimodal perception framework.

disparity, computing left and right disparity maps and performing left–right consistency validation (which in our adaptation is used to produce the stereovision confidence maps).

- *Binaural sensor system.* Using an adaptation of the real-time software kindly made available by the Speech and Hearing Group at the University of Sheffield (Lu, Christensen, & Cooke, 2007) to

implement binaural cue analysis as described in Pinho et al. (2008) and Ferreira et al. (2008b).
- *Bayesian Volumetric Map, Bayesian sensor models and active exploration*. Using our proprietary, parallel processing, GPU implementation developed with NVIDIA's general purpose parallel computing architecture CUDA (NVIDIA, 2007).

Visual saliency properties

Saliency properties from a generic visual cue, or, in other words, the conspicuity maps given by the BVM extended operators $Q_C^{i,t} = P([S_C^{i,t} = 1]|Z_j^{i,t}\, C) \in [0, 1]$, were implemented in two steps.

1. A single-channel image with values varying between 0 and 1 is taken directly from visual cues taken from the right camera of the stereovision setup (thus simulating a dominant eye), either by directly normalizing traditional dense conspicuity maps as defined by Itti et al. (1998), or by generating a conspicuity map by forming Gaussian distributions with specific standard deviations centred on individual points of interest on the right camera image, for example in the case of sparse feature extractors such as face detection algorithms.
2. The saliency values from each pixel in the conspicuity map for which a disparity was estimated by the stereovision module are then projected on the log-spherical configuration through projection lines spanning the corresponding (θ, ϕ) angles – if two or more different saliency values are projected throughout the same direction, only the highest saliency value is used. These values are thus considered as soft evidence regarding $S_C^{i,t}$, therefore yielding $Q_C^{i,t}$.

The specific properties used in this work (although any visual saliency property would have been usable by applying the two steps described above) were optical flow magnitude taken from the result of using the CUDA implementation of the 'Bayesian Multi-scale Differential Optical Flow' algorithm of Simoncelli (1999) by Hauagge (2009), and face detection using the Haar-like features implementation of the OpenCV library. Using these implementations, the 15 Hz performance of the stereovision unit where they are integrated, as reported in Ferreira et al. (2011) and Ferreira et al. (2009), is reduced to about 6 Hz, mainly as a consequence of the slow performance of the face detection algorithm.

Auditory saliency properties

The auditory saliency property used in this work was directly implemented from the $P([S_C = 1]|Z\,C)$ question solved by the Bayesian model of binaural perception.

Inhibition of Return

The IoR mechanism used in this work is implemented by assigning values on a log-spherical data structure corresponding to R_C^t ranging from 1 to values close to 0 depending on the distance in $\mathcal{Y}$ between G^{t-1} and each C, denoted d_{IoR}, through the following expression

$$R_C^t \equiv f(G^{t-1}) = \left(\frac{1}{2}\right)^{d_{IoR}} \tag{5}$$

Hierarchical model

The parameters of the Beta distributions defined on the Bayesian Programs of Figures 4, 5 and 6 will, in general, function as relative importance weights for each behaviour in the fusion process. However, in two extreme cases, theses parameters will serve as a switch: $\alpha = 1, \beta = 1$ will result in degenerating the corresponding beta distribution into a uniform distribution, hence switching off the respective behaviour, while $\alpha >> \beta$ or $\beta >> \alpha$ will degenerate the Beta distribution into a Dirac delta function, hence serving as a mutually exclusive switch, 'numerically deactivating' all other behaviours.

A set of parameters was chosen for initial values in order to attain the beta distributions presented in Figure 9. These preprogrammed parameters define the genetic imprint of preliminary knowledge that establishes the baseline hierarchy of the set of active perception behaviours; these parameters are changeable 'on the fly' through sliders on the graphical user interface of the implementation software, thus simulating top-down influences on behaviour prioritization (i.e. the *adaptivity* property). The influence of the relative weights imposed by this choice of parameters will be discussed in the Results section.

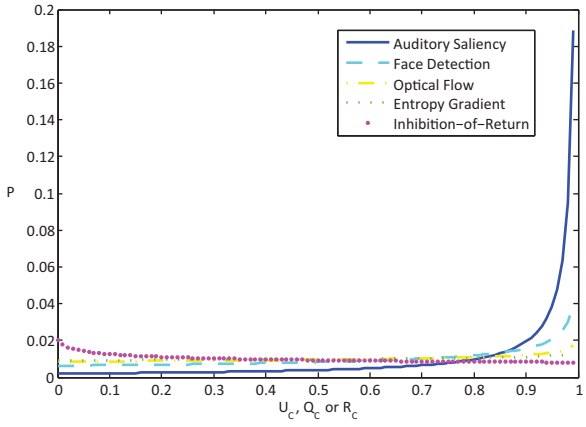

Figure 9. Beta distributions of the active perception hierarchy using the baseline choice for parameters. Corresponding parameters are $\alpha_U = 1$ and $\beta_U = 0.92$ for active exploration, $\alpha_Q = 1$ and $\beta_Q = 0.01$ for auditory saliency, $\alpha_Q = 1$ and $\beta_Q = 0.6$ for face detection saliency, $\alpha_Q = 1$ and $\beta_Q = 0.85$ for optical flow magnitude saliency, and $\alpha_R = 0.8$ and $\beta_R = 1$ for IoR.

The fixation point for the next time instant G^t is obtained by substituting equations (2) and (3) consecutively into (4), and computing $G^t = \text{argmax}_C P([G^t = C]|V^{1\rightarrow t} O^{1\rightarrow t} S^t G^{t-1} \pi_C)$, knowing that

$$P([G^t = C]|V^{1\rightarrow t} O^{1\rightarrow t} S^t G^{t-1} \pi_C) \propto P(U_C^t|[G^t = C] \pi_A) \prod_{i=1}^{N} \left[P(Q_C^{i,t}|[G^t = C] \pi_B)\right] P(R_C^t|[G^t = C] \pi_C) \quad (6)$$

by factoring out the effect of the uniform distributions corresponding to considering $[G^t \neq C]$.

Finally, the full active perception system runs at about 5 Hz, for $N = 10$, $\Delta\theta = 1°$, $\Delta\phi = 2°$, mainly due to the degraded performance of the stereovision unit reported above. In any case, these ratings are still just within the parameters of satisfactory real-time performance, as defined in Ferreira et al. (2011).

Results and discussion

Five experimental sessions were conducted to test the performance of the hierarchical framework presented in this text, in particular to demonstrate its properties of emergence, scalability and adaptivity, as summarized in Table 1. Several repetitions of each of these sessions were conducted under roughly the same conditions, so as to confirm reproducibility of the same behaviours. Consequently, in the following lines, the results of each of these sessions will be discussed, and a summary of these findings will be presented in Table 2.

During all the experiments, three views were also filmed from external cameras – see Figure 10 for an overview of the experimental setup using one of these views – and a body-tracking suit was also used by the speaker to the left from the IMPEP head's perspective, the only speaker allowed to walk from one position to another within the BVM horopter (i.e. the portion of spherical volume being perceived and consequently represented by the map), for positioning ground-truth.

Experimental Session 1 – active perception hierarchy implementing all behaviours, using baseline priorities

In this session, a two-speaker scenario was enacted following a script (Figure 11) roughly describing the activity reported in the annotated timeline of the experiment presented in Figure 12 (a video of Session 1 is available as supplementary online material).

The genetically imprinted parameters for the distributions that was used was presented on Figure 9. This particular choice of parameters was made to emphasize socially-oriented, high-level behaviours as opposed to low-level behaviours and the IoR effect, which has a noticeable effect only when the former are absent. Countering the IoR effect in the presence of socially-oriented behaviours allows for an apparently more natural emergent behaviour of the system.

The empirical process of finding working parameters within the restrictions described above involved minimal trial-and-error. The greatest restriction was found to be the proportion between weights of auditory saliency and of visual saliency (more specifically, in this case, face detection, the second highest priority). Nevertheless, the range of acceptable proportions was still found to be large enough to be easy to pinpoint.

As can be seen in Figure 12, the system successfully fixated both speakers, and even exhibited an emergent behaviour very similar to smooth pursuit while following the speaker to the left in the perspective of the

Table 1. Summary table of experimental session planification.

Session	Description	Objective
1	Implement all behaviours, using baseline priorities	Baseline demonstration, proof-of concept of emergence
2	Implement all behaviours, using swapped priorities	Proof-of-concept of adaptivity
3	Implement active exploration only	1st proof-of concept of scalability
4	Implement optical flow magnitude saliency only	2nd proof-of-concept of scalability
5	Implement optical flow IoR only	3rd proof-of-concept of scalability

Table 2. Summary table of experimental session results.

Session	Result
1	System performance was annotated by human evaluators, and logs of saliency maps were analysed in comparison, showing that an appropriate choice of weights results in a reasonably human-like emergent behaviour, essential for HRI.
2	System performance was evaluated, as with session 1, proving that modifying the weights for each behaviour change emergence drastically, thus demonstrating the framework's adaptivity.
3, 4 and 5	Removing a specific basic behaviour is shown to change emergent behaviour while maintaining consistency, thus demonstrating the scalability of the system.

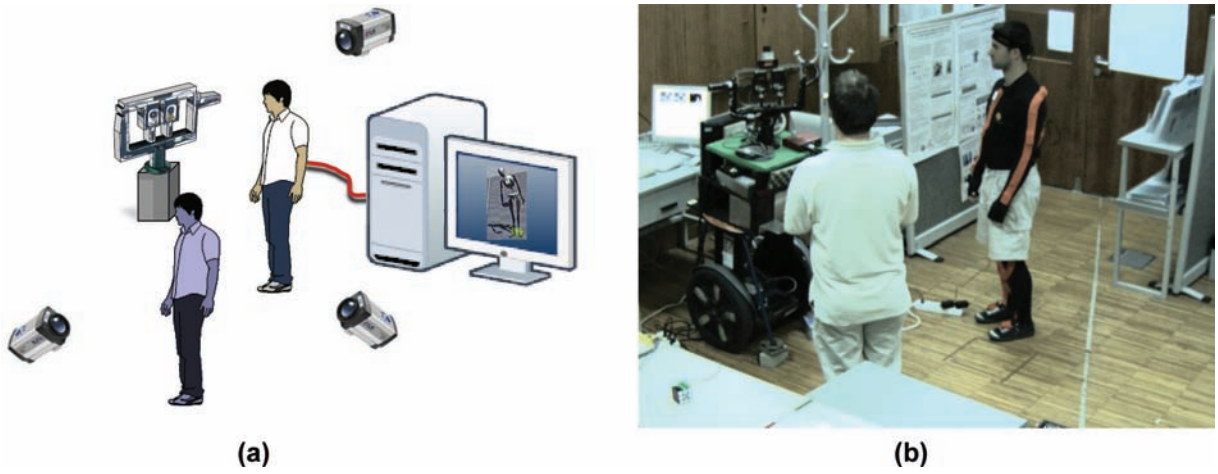

Figure 10. Overview of the setup used in the experimental sessions testing the Bayesian hierarchical framework for multimodal active perception. The 'IMPEP 2 and interlocutors' scenario, in which one of the interlocutors is wearing body-tracking suit, is implemented using an acting script (presented on Figure 11). During the experimental sessions, the signals which were recorded for analysis included data from: IMPEP 2 time-stamped video and audio logging; camera network capturing several external points of view; body-tracking poses. All signals were synchronized through common-server timestamping.

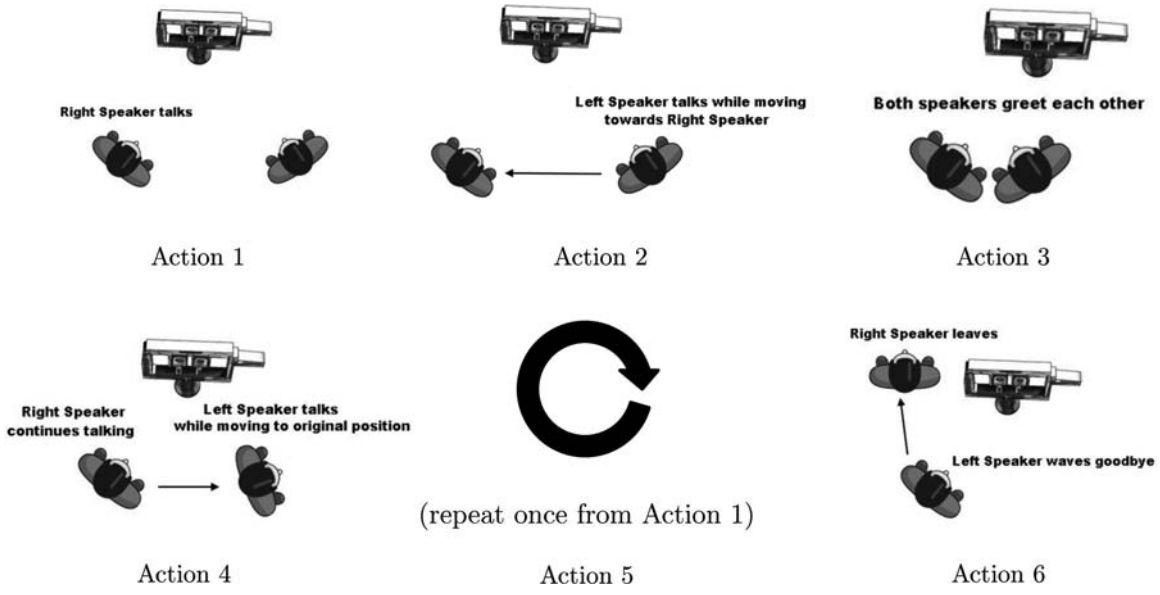

Figure 11. Acting script for active perception experiments.

IMPEP head. Therefore, a purely saccade-generating model yields, through emergence, a behaviour that closely resembles smooth pursuit. After analysing logs for $P([G^t = C]|V^{1\to t}O^{1\to t}S^t G^{t-1}\pi_C)$, it was found that probabilities for saliency moved across the occupancy grid smoothly, given system parameters and temporal performance. This shows that the baseline priority rationale for the choice of parameters for the distributions was reasonably planned, but, more importantly, clearly demonstrates emergence due to fusion as more than just a pure 'sum of parts'.

Offline high-definition renderings of BVM and saliency logs are presented on Figures 13 and 14, respectively.

Experimental Session 2 – active perception hierarchy implementing all behaviours, with swapped priorities

In this session, the first part of the script of Experimental Session 1 was reenacted, but this time swapping the parameters of the distributions for auditory saliency and face detection saliency, presented on Figure 9. This resulted in the system being unable to change gaze direction to the second speaker after fixating the first speaker, due to the deadlock caused by the face detection saliency keeping attention on the first speaker's face, further showcasing the importance of choosing the appropriate weights for each behaviour.

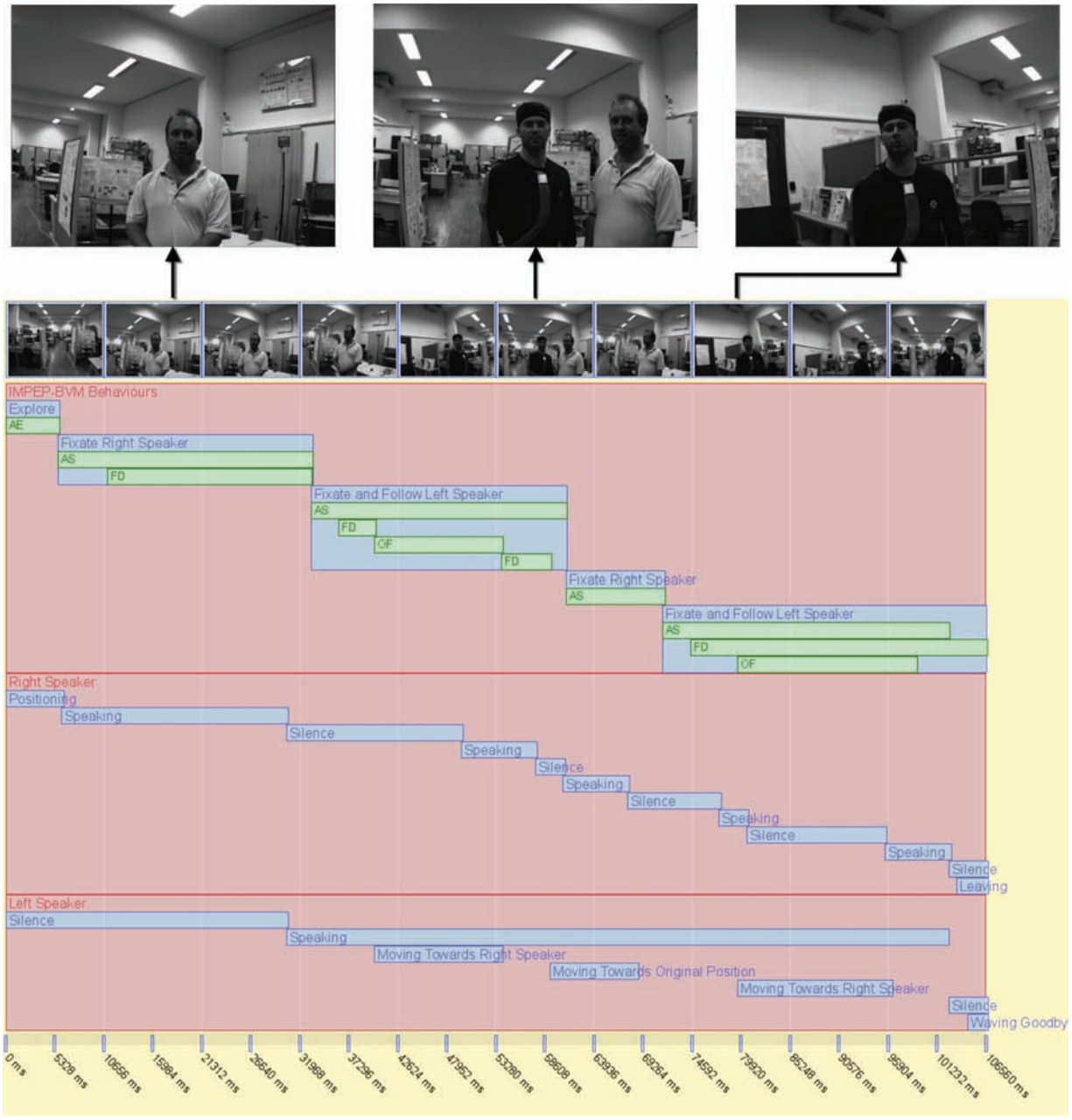

Figure 12. Annotated timeline for Experimental Session 1 – active perception hierarchy implementing all behaviours using baseline priorities. Annotation was performed by three different observers, who were naive to the underlying basic behaviours and weighting process; the timeline shows the rough average of these annotations, which were very similar (subtle semantic differences between annotation label texts were filtered out, resulting in a maximum temporal difference of annotations of approximately 1s between annotators). The two lower annotation lanes, labelling the actions performed by the right and left speaker in the perspective of the IMPEP head, were performed by inspection of images taken by the IMPEP stereovision system, by the external cameras, by the tracking suit, and by the audio file recorded by the IMPEP binaural system. The top annotation lane, labelling the emergent behaviours of the active perception system and an interpretation of what were the most prominent underlying low-level behaviours (AE: active exploration; AS: auditory saliency; OF: optical flow magnitude saliency; FD: face detection saliency), was annotated by additionally inspecting saved logs of $P([G^t = C]|V^{1\to t}O^{1\to t}S^t G^{t-1}\pi_C)$.

Experimental Session 3 – active perception hierarchy implementing active exploration only

In this session, the full script of Experimental Session 1 was reenacted, but this time all behaviours except entropy gradient-based active exploration were turned off by making all other distributions uniform. As expected, the behaviour described in Ferreira et al. (2009) emerged, namely the typical 'chicken-like' saccadic movements of the IMPEP head exploring the surrounding environment, and a particular sensitivity to the entropy caused by binaural sensing and motion.

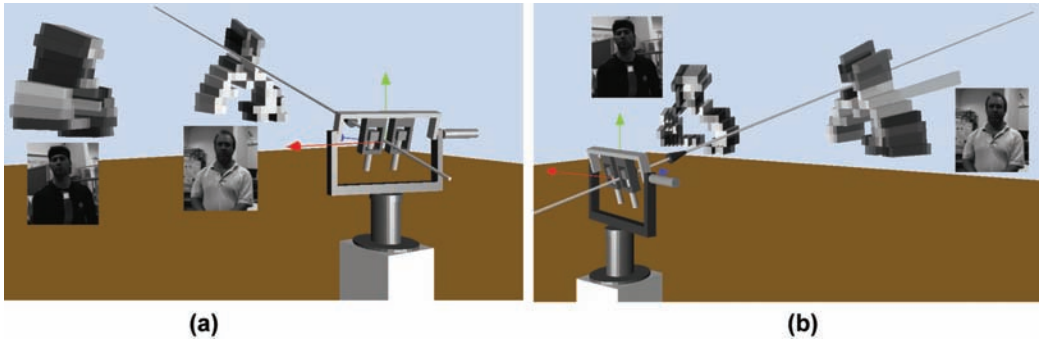

Figure 13. Offline rendering of a BVM representation of the two speakers scenario of Experimental Session 1. After the experiment, an instantiation of the BVM occupancy grid is rendered using a Blender-based viewer, of which two different views are presented. Notice the well-defined speaker upper torso silhouette reconstructions, which are clearly identifiable even despite the distortion elicited to visual inspection caused by the log-spherical nature of each cell. The oriented 3D sketch of the IMPEP perception system denotes the current gaze orientation. All results depict frontal views, with Z pointing outward. The parameters for the BVM are as follows: $N = 10$, $[\rho_{Min} = 1000]$mm and $[\rho_{Max} = 2500]$mm, $\theta \in [-180°, 180°]$, with $\Delta\theta = 1°$, and $\phi \in [-90°, 90°]$, with $\Delta\phi = 1°$, corresponding to $10 \times 360 \times 180 = 648{,}000$ cells, approximately delimiting the so-called 'personal space' (the zone immediately surrounding the observer's head, generally within arm's reach and slightly beyond, within 2 m range (Cutting & Vishton, 1995)).

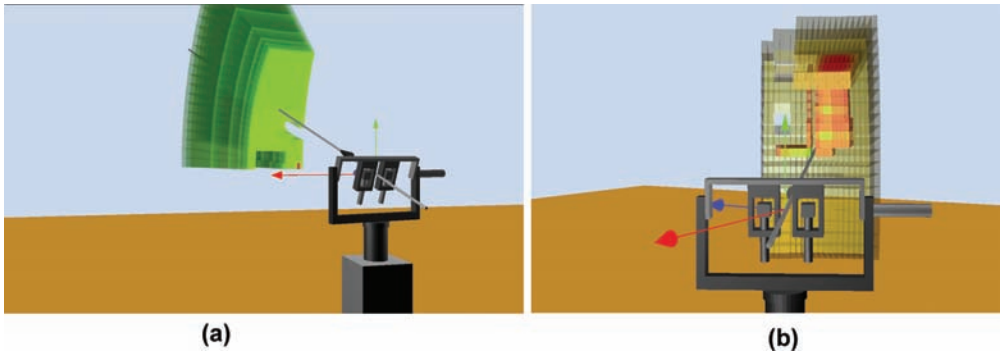

Figure 14. Offline rendering of example saliency maps of the two speakers scenario of Experimental Session 1. The rendering represents values for $P([G^t = C]|V^{l \to t} O^{l \to t} S^t G^{t-1} \pi_C)$ that were logged during the session for a specific time instant. Only a slice corresponding to all cells at 10° in azimuth and 20° in elevation around the next fixation point G^t with $P(O_C|C) > .45$ are shown, depicted using a smoothly gradated light-to-dark colour-code (dark corresponds to lower values, light corresponds to higher values). All other parameters and labelling are the same or analogous to Fig 13. On the left, a purely auditory-elicited map is shown, while on the right, a map resulting from the fusion of at least auditory and face detection conspicuity maps is shown.

Experimental Session 4 – active perception hierarchy implementing optical flow magnitude saliency only

In this session, a single human subject (using the body-tracking suit) is tracked while walking from one position to another within the system's horopter using only optical flow magnitude saliency by making all other distributions uniform, as before. As long as the subject walked within reasonable velocity limits, the system was able to track them successfully.

A saliency map from this session, representing an example of an optical flow magnitude conspicuity map, is presented on Figure 15.

Experimental Session 5 – active perception hierarchy implementing Inhibition of Return only

In this session, the IoR behaviour was tested by making all other distributions uniform, as before. In this case, a fortuitous saccadic behaviour emerged, with the system redirecting gaze to random directions at a constant rate.

A summary of the experimental session results can be found in Table 2. AQ3

Overall conclusions and future work

In conclusion, the Bayesian hierarchical framework presented in this article was shown to adequately follow human-like active perception behaviours, namely by exhibiting the following desirable properties.

Emergence. High-level behaviour results from low-level interaction of simpler building blocks.

Scalability. Seamless integration of additional inputs is allowed by the Bayesian Programming formalism used to state the models of the framework.

Sdaptivity. The initial 'genetic imprint' of distribution parameters may be changed 'on the fly' through

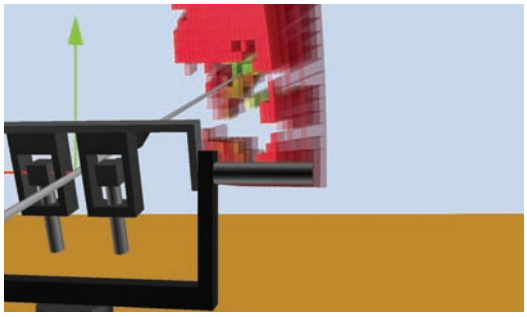

Figure 15. Offline rendering of an example optical flow magnitude saliency map of Experimental Session 4. All parameters and labelling are the same or analogous to Fig 13.

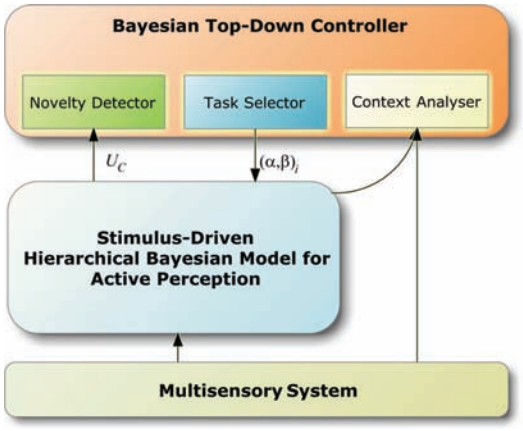

Figure 16. Proposal for goal-oriented active perception framework, including both bottom-up and top-down influences. Corbetta & Shulman (2002) posited the hypothesis that a part of the human brain's ventral system acts as a circuit breaker of ongoing cognitive activity when a behaviourally relevant stimulus is detected. According to these authors, when subjects detect an unexpected event, they must break the current attentional set and adopt a new one on the basis of the incoming stimulus. The framework presented in this diagram implements this hypothesis, assuming the set of parameters (α,β) as the 'current attentional set', as defined by Corbetta & Shulman (2002), with the entropy gradient-based factor U_C for the current time instant being checked for abrupt changes in order for the novelty detector to recognize unexpected events.

parameter manipulation, thus allowing for the implementation of goal-dependent behaviours (i.e. top-down influences).

Future improvements to this framework naturally involve taking advantage of its scalability to include new relevant behaviours and, most importantly, to capitalize on its adaptivity in order to implement a goal-oriented system in which active perception emergent behaviour changes depending on the task being performed by the robot (Figure 16).

Further future work involves exploiting an important facet of Bayesian frameworks, namely parameter learning – the system will be trained by human subjects using a head-mounted device. The subjects' tracked head–eye gaze shifts control the virtual stereoscopic-binaural point of view, and hence the progression of each stimulus movie – see Figure 17 – while logs of audiovisual stimuli and corresponding fixation points will be logged. This way, controlled free-viewing conditions will be enforced by proposing both generic and specific tasks to the subjects, thus enabling a systematic estimation of the distribution parameters in order to construct a lookup table of weights to promote the appropriate human-like emergent behaviour depending on the robot's goal. On the other hand, this learning process will allow testing both of our primary hypotheses for active visuoauditory perception, namely active exploration and automatic orienting using sensory saliency, as valid strategies in human behaviour regarding saccade generation.

More details and results concerning the work presented herewith can be found at http://paloma.isr.uc.pt/jfilipe/BayesianMultimodalPerception.

Notes

1. Positions in space could be described in either of the two camera coordinate systems, but it is convenient to describe positions relative to an imaginary third camera half-way between the two real cameras. This is commonly called the *cyclopean coordinate system* (the cyclops only had one eye in the middle of its head). The percept of a single image elicited by the human brain, when stereoscopic fusion is correctly performed, is, in fact, a consequence of the application of this geometry.

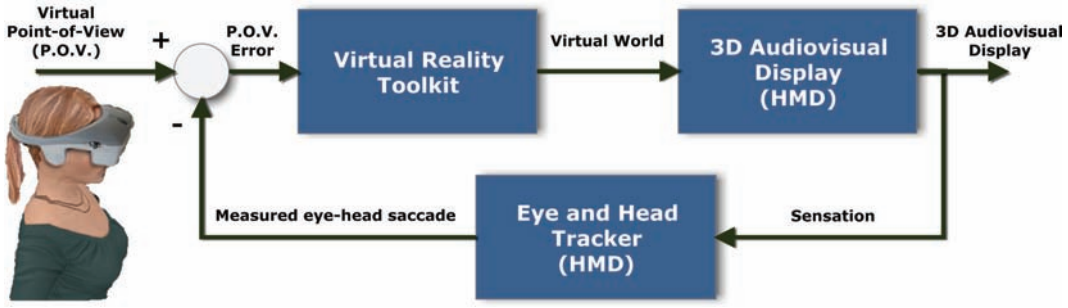

Figure 17. Virtual point-of-view generator setup that allows the updating of audiovisual stimuli presentation according to the monitored subjects' gaze direction.

2. *Soft evidence* is a concept of Bayesian theory, used, for example, in Jeffrey's rule – an auxiliary random variable represents the probability of evidence of another variable, thus providing an indirect means of accessing that evidence. For more information, please refer to Pearl (1988).

Acknowledgements

The authors would particularly like to thank, at the Institute of Systems and Robotics (ISR/FCT-UC), Pedro Trindade for developing the Blender-based BVM viewer software and João Quintas for his inestimable help with the concluding experimental work and, at the Institute of Biomedical Research in Light and Image (IBILI/UC), Gil Cunha and João Castelhano for their help with the future work section.

Funding

This research has been supported by the Portuguese Foundation for Science and Technology (FCT) [post-doctoral grant number SFRH/BPD/74803/2010].

References

Aloimonos, J., Weiss, I., & Bandyopadhyay, A. (1987). Active vision. *International Journal of Computer Vision*, *1*, 333–356.

Bajcsy, R. (1985). Active perception vs passive perception. In *Third IEEE Workshop on Computer Vision* (pp. 55–59), Bellaire, Michigan.

Ballard, D. H. (1999). *An introduction to natural computation*. Cambridge, MA: MIT Press.

Bessière, P., Laugier, C., & Siegwart, R. (Eds.). (2008). *Probabilistic reasoning and decision making in sensory-motor systems* (Vol. 46). Berlin: Springer.

Bohg, J., Barck-holst, C., Huebner, K., Ralph, M., Rasolzadeh, B., Song, D., et al. (2009). Towards grasp-oriented visual perception for humanoid robots. *International Journal of Humanoid Robotics, 3*(3), 387–434.

Breazeal, C., Edsinger, A., Fitzpatrick, P., & Scassellati, B. (2001). Active vision for sociable robots. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and* Humans, *31*(5), 443–453.

Buntine, W. L. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research (AI Access Foundation), 2*, 159–225.

Carmi, R., & Itti, L. (2006). Causal saliency effects during natural vision. In *ACM eye tracking research and applications* (pp. 1–9).

Colas, F., Diard, J., & Bessière, P. (2010). Common Bayesian models for common cognitive issues. *Acta Biotheoretica, 58*(2–3), 191–216.

Colas, F., Flacher, F., Tanner, T., Bessière, P., & Girard, B. (2009). Bayesian models of eye movement selection with retinotopic maps. *Biological Cybernetics, 100*, 203–214.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*, 201–215.

Croon, G. C. H. E. de, Sprinkhuizen-Kuyper, I. G., & Postma, E. O. (2009). Comparing active vision models. *Image and Vision Computing, 27*(4), 374–384.

Cutting, J. E., & Vishton, P. M. (1995). Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In: W. Epstein & S. Rogers (Eds.), *Handbook of perception and cognition* (Vol. 5; Perception of space and motion). Academic Press.

Dankers, A., Barnes, N., & Zelinsky, A. (2005). Active vision for road scene awareness. In *IEEE Intelligent Vehicles Symposium (IVS05)* (pp. 187–192), Las Vegas, USA.

Dankers, A., Barnes, N., & Zelinsky, A. (2007). A Reactive vision system: Active-dynamic saliency. In *5th International Conference on Computer Vision Systems*, Bielefeld, Germany.

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (Eds.). (2007). *Bayesian brain — probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.

Elfes, A. (1989). Using occupancy grids for mobile robot perception and navigation. *IEEE Computer, 22*(6), 46–57.

Ferreira, J. F., Bessière, P., Mekhnacha, K., Lobo, J., Dias, J., & Laugier, C. (2008). Bayesian models for multimodal perception of 3D structure and motion. In *International Conference on Cognitive Systems (CogSys 2008)* (pp. 103–108), University of Karlsruhe, Karlsruhe, Germany.

Ferreira, J. F., & Castelo-Branco, M. (2007). *3D structure and motion multimodal perception* (State-of-the-Art Report). Institute of Systems and Robotics and Institute of Biomedical Research in Light and Image, University of Coimbra. (Bayesian Approach to Cognitive Systems (BACS) European Project)

Ferreira, J. F., Lobo, J., & Dias, J. (2011). Bayesian real-time perception algorithms on GPU — Real-time implementation of Bayesian models for multimodal perception using CUDA. *Journal of Real-Time Image Processing*, *6*(3), 171–186.

Ferreira, J. F., Pinho, C., & Dias, J. (2008a). Active exploration using Bayesian models for multimodal perception. In A. Campilho & M. Kamel (Eds.), *Image Analysis and Recognition, Lecture Notes in Computer Science series, International Conference ICIAR 2008* (pp. 369–378). Berlin: Springer.

Ferreira, J. F., Pinho, C., & Dias, J. (2008b). Bayesian sensor model for egocentric stereovision. In *14ª Conferência Portuguesa de Reconhecimento de Padrões Coimbra (RECPAD 2008)*.

Ferreira, J. F., Pinho, C., & Dias, J. (2009). Implementation and calibration of a Bayesian binaural system for 3D localisation. In *2008 IEEE International Conference on Robotics and Biomimetics (ROBIO 2008)*, Bangkok, Thailand.

Ferreira, J. F., Prado, J., Lobo, J., & Dias, J. (2009). Multimodal active exploration using a Bayesian approach. In *IASTED International Conference in Robotics and Applications,* (pp. 319–326), Cambridge MA, USA.

Gallup, D. (2009). *CUDA Stereo*. Available from: http://www.cs.unc.edu/~gallup/stereo-demo.

Hauagge, D. C. (2009). *CUDABOF — Bayesian Optical Flow on NVidia's CUDA*. Available from http://www.liv.ic.unicamp.br/~hauagge/Daniel_Cabrini_Hauagge/Home_Page.html

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences, 27*(12), 712–719.

Koene, A., Morén, J., Trifa, V., & Cheng, G. (2007). Gaze shift reflex in a humanoid active vision system. In *5th International Conference on Computer Vision Systems*, Bielefeld, Germany.

Kopp, L., & Gärdenfors, P. (2002). Attention as a minimal criterion of intentionality in robots. *Cognitive Science Quarterly, 2*, 302–319.

Lebeltel, O. (1999). *Programmation Bayésienne des Robots*. Unpublished doctoral dissertation, Institut National Polytechnique de Grenoble, Grenoble, France.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology, 55*(1), 1–7. (Special Issue on Hierarchical Bayesian Models)

Lu, Y.-C., Christensen, H., & Cooke, M. (2007). Active binaural distance estimation for dynamic sources. In *Interspeech 2007* (pp. 574–577), Antwerp, Belgium.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. S. Francisco: W. H. Freeman and Company.

Niebur, E., Itti, L., & Koch, C. (1995). Modeling the ''where'' visual pathway. In T. J. Sejnowski (Ed.), *2nd Joint Symposium on Neural Computation, Caltech-UCSD* (Vol. 5, pp. 26–35), La Jolla.

NVIDIA. (2007). *CUDA Programming Guide version 1.2*.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference* (Revised Second Printing ed.; M. B. Morgan, Ed.). Morgan Kaufmann Publishers, Inc. (Elsevier).

Pinho, C., Ferreira, J. F., Bessière, P., & Dias, J. (2008, April). A Bayesian binaural system for 3D sound-source localisation. In *International Conference on Cognitive Systems (CogSys 2008)* (p. 109–114), University of Karlsruhe, Karlsruhe, Germany.

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience, 1*, 125–132. (Review)

Shibata, T., Vijayakumar, S., Conradt, J., & Schaal, S. (2001). Biomimetic oculomotor control. *Adaptive Behaviour - Special Issue on Biologically Inspired and Biomimetic System, 9*(3–4), 189–208.

Shiffrin, R., Lee, M., Wagenmakers, E.-J., & Kim, W. J. (2008). A survey of model evaluation approaches with a focus on hierarchical Bayesian methods. *Cognitive Science, 32*(8), 1248–1284.

Simoncelli, E. P. (1999). Bayesian multi-scale differential optical flow. In B. Jähne, H. Haussecker, & P. Geissler (Eds.), *Handbook of Computer Vision and Applications*. Academic Press.

Tay, C., Mekhnacha, K., Chen, C., Yguel, M., & Laugier, C. (2008). An efficient formulation of the Bayesian occupation filter for target tracking in dynamic environments. *International Journal of Autonomous Vehicles, 6*(1–2), 155–171.

Tsotsos, J., & Shubina, K. (2007). Attention and visual search: Active robotic vision systems that search. In *The 5th International Conference on Computer Vision Systems*, March 21–24, Bielefeld.

João Filipe Ferreira is currently an Invited Assistant Professor at the Department of Electrical Engineering and Computer Science, Faculty of Sciences and Technologies of the University of Coimbra, and a Post-Doc at the Institute of Systems and Robotics (ISR), Coimbra Pole, working on the subject "Multimodal Active Perception Framework Using a Bayesian Approach", sponsored by a scholarship from the national Foundation for Technology and Sciences (FCT).

He received his Ph.D. in Electrical Engineering from the University of Coimbra, specialisation in Instrumentation and Control, in July 2011, on the subject "Bayesian Cognitive Models for 3D Structure and Motion Multimodal Perception". He received his Electrical Engineering degree (B.Sc., specialisation in computers) from the Faculty of Sciences and Technology, University of Coimbra (FCTUC) in July 2000. He received the M.Sc. degree in Electrical Engineering from the same faculty, specialisation in Automation and Robotics, in January 2005.

His main research interests are Bayesian approaches to modelling, human and artificial perception, robotics, image processing and 3D scanning.

He conducts his research on Bayesian cognitive models for multimodal artificial perception systems at the Electrical Engineering Department of the same Faculty, and research on human multimodal perception at the Biomedical Institute for Research in Light and Image (IBILI), Faculty of Medicine of the University of Coimbra. He is also a staff researcher at the ISR since 1999.

He was a staff researcher for the FCTUC team on the European Integrated Project "Bayesian Approach to Cognitive Systems" (FP6-IST-27140), from 2006 to 2010, and a research fellow for the European Integrated Project HANDLE "Developmental pathway towards autonomy and dexterity in robot in-hand manipulation" (theme 2: Cognitive Systems, Interaction, Robotics, under grant agreement 231640) from 2010 to 2011.

Miguel Castelo-Branco is currently an Assistant Professor at the University of Coimbra, Portugal, and has held a similar position in 2000 at the University of Maastricht, the Netherlands. Before (1998-1999), he was a Postdoctoral fellow at the Max-Planck-Institut für Hirnforschung, Germany where he had also performed his PhD work (1994-1998). He is also

the Director of IBILI (Institute for Biomedical Research on Light and Image), Faculty of Medicine, Coimbra, Portugal.

He has made contributions in the fields of Ophthalmology, Neurology, Visual Neuroscience, Human Psychophysics, Functional Brain Imaging and Human and Animal Neurophysiology. His lab is now accomplishing tasks also in the context of a European Network (Evi-Genoret), and has succeed in collaborating with labs working in other fields of knowledge such as Human Genetics and Clinical Neuroscience. He is also involved in the National Functional Brain Imaging Network. In his work he could isolate specific magnocelullar/visual motion dysfunction in a genetic neurodevelopmental model, Williams Syndrome. He has further studied parallel pathways to quantitatively analyze visual aging in neurodegenerative disorders of the retina and the brain (Glaucoma and Parkinson Disease). His laboratory is very experienced in Visual Impairment questions, and the multiple causes of amblyopia and its functional characterization in centre and peripheral visual field. In recent work, the lab has characterized genetic and acquired photoreceptor retinal degenerations. One major goal is to provide models of visual impairment based on new structure-function and genotype-phenotype correlations (that may help define new biomarkers for retinal degenerations). He has also published work on neuropsychology and psychophysics in patient populations. His achievements are well reflected in publications in top General Journals, such as Nature and PNAS and Top Translational research journals such as Journal of Clinical Investigation (impact factor(IF): 17), Brain (IF 8) as well as others in the field of Basic and Clinical Visual Sciences.

Jorge Manuel Miranda Dias received his Ph.D. on Electrical Engineering by the University of Coimbra, specialization in Control and Instrumentation, in November 1994. He holds his research activities on the System and Robotics Institute (Instituto de Sistemas e Robótica) from the University of Coimbra. Jorge Dias research area is Computer Vision and Robotics, with activities and contributions on the field since 1984. He has several publications on Scientific Reports, Conferences, Journals and Book Chapters. Jorge Dias teaches several engineering courses at the Electrical Engineering and Computer Science Department- from the Faculty of Science and Technology - University of Coimbra. He is responsible by courses on Computer Vision, Robotics, Industrial Automation, Microprocessors and Digital Systems. He is also responsible by the supervision of Master and Ph.D. students on the field of Computer Vision and Robotics. He was main researcher from projects financed by European Community (Framework Programe 6 and 7), by the Portuguese Foundation for Science and Technology. He is Senior Member of IEEE - the world's largest professional association for the advancement of technology. He is currently the officer in charge for the Portuguese Chapter for IEEE – RAS (Robotics and Automation Society), and also the vice-president of "Sociedade Portuguesa de Robótica - SPR".