Cogn Process
DOI 10.1007/s10339-012-0481-9

SHORT REPORT

Learning emergent behaviours for a hierarchical Bayesian framework for active robotic perception

João Filipe Ferreira · Christiana Tsiourti · Jorge Dias

Abstract In this research work, we contribute with a behaviour learning process for a hierarchical Bayesian framework for multimodal active perception, devised to be emergent, scalable and adaptive. This framework is composed by models built upon a common spatial configuration for encoding perception and action that is naturally fitting for the integration of readings from multiple sensors, using a Bayesian approach devised in previous work. The proposed learning process is shown to reproduce goal-dependent human-like active perception behaviours by learning model parameters (referred to as "attentional sets") for different free-viewing and active search tasks. Learning was performed by presenting several 3D audiovisual virtual scenarios using a head-mounted display, while logging the spatial distribution of fixations of the subject (in 2D, on left and right images, and in 3D space), data which are consequently used as the training set for the framework. As a consequence, the hierarchical Bayesian framework adequately implements high-level behaviour resulting from low-level interaction of simpler building blocks by using the attentional sets learned for each task, and is able to change these attentional sets "on the fly," allowing the implementation of goal-dependent behaviours (i.e., top-down influences).

J. F. Ferreira (✉) · C. Tsiourti · J. Dias
ISR, University of Coimbra, Coimbra, Portugal
e-mail: jfilipe@isr.uc.pt

J. Dias
Khalifa University of Science, Technology and Research (KUSTAR), Abu Dhabi, UAE

Keywords Multisensory active perception · Hierarchical Bayes models · Bioinspired robotics · Human–robot interaction · Emergence · Scalability · Adaptive behaviour

Introduction

How should the uncertainty and incompleteness of the environment be represented and modelled so as to increase the autonomy of a robot? Can a robotic system perceive, infer, decide and act more efficiently by using a probabilistic framework? These are two of the challenging questions robotics researchers are currently facing in the design of more autonomous and intelligent artificial robotic systems.

Previous work conducted in the Institute of Systems and Robotics by Ferreira et al. (2012, 2011)—see also (Ferreira 2011)—has contributed to addressing these challenges by providing the basis of a framework for artificial active multimodal perception, comprised of a real-time GPU-based implementation of a scalable, adaptive and emergent hierarchical Bayesian active perception system that simulates several bottom-up-driven human behaviours of attention guidance—see Fig. 1. It was devised mainly to be used in human–robot interaction (HRI) applications.

These emergent behaviours are implemented by combining simple behaviours using a set of weights, thus implementing a process analogous to the attentional set as defined by Corbetta and Shulman (2002). The bottom layer of this framework consists of a log-spherical inference grid updated using a Bayesian filter, the Bayesian volumetric map or BVM. Ferreira et al. (2012, 2011) modelled visuoauditory perception using an approach that finds its inspiration in the fast pathways believed to exist in the human brain, which are closely linked to primal instincts of survival and basic social interaction.

Journal : Large 10339 | Dispatch : 2-7-2012 | Pages : 5
Article No. : 481 | LE | TYPESET
MS Code : | CP | DISK

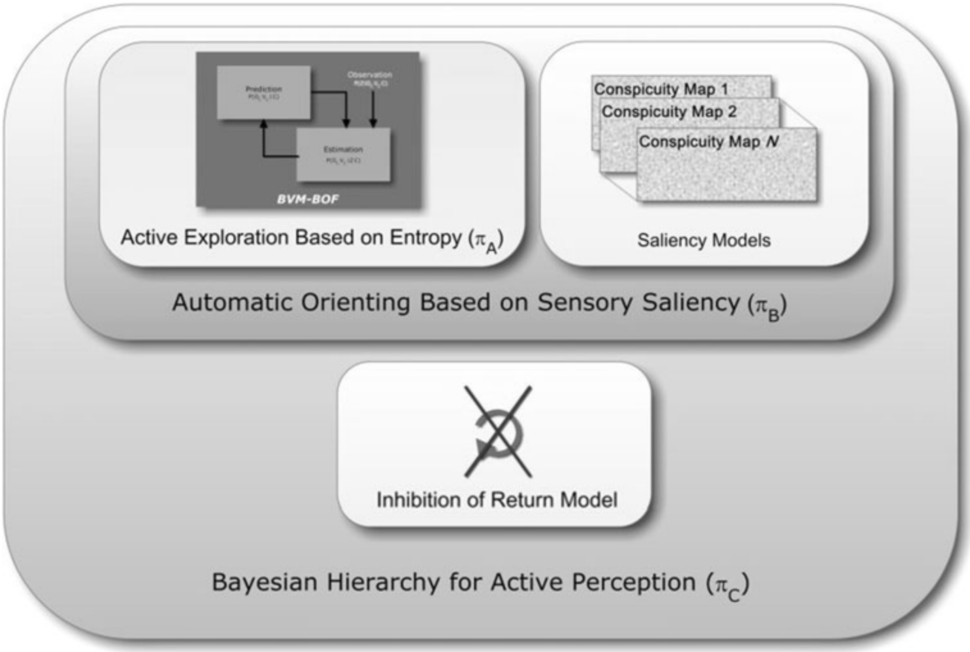

Fig. 1 Conceptual diagram for active perception model hierarchy (Ferreira et al. 2012)

In this text, we will give a brief overview of a behaviour learning process for this framework, designed to estimate its free parameters (identified as attentional sets) for different free-viewing and active search tasks.

Related work

During the last decades, researchers from different fields (psychologists, neuroscientists and, more recently, computer scientists) have investigated visual attention thoroughly and also, to a lesser extent, visuoauditory attention, both in terms of model analysis and in terms of synthesis and implementation.

In the fields of neuroscience and psychology, for example, research on these issues has ranged from the early twentieth century, such as the work by Buswell (1935), to recently Castelhano et al. (2009), and separately and subsequently also Mills et al. (2011), which investigated the influence of task instruction on specific parameters of eye movement control, such as the number of fixations and gaze duration on specific objects. On the other hand, research work in computational models of artificial active perception has ranged from the seminal work of Bajcsy (1985) and Aloimonos et al. (1987), through Itti et al. (1998) in the bottom-up influence of visual saliency and Breazeal et al. (2001) in active vision for social robots, to important and recent work that has attempted to implement learning by imitation for active perception behaviours, such as the work by Belardinelli et al. (2007).

We improve the current state of the art in all of these fields by contributing in two specific fronts: (1) a learning process integrated within psychophysical experiments, representing an important tool for both model analysis and synthesis in human studies and robot development, respectively; and (2) a fully integrated probabilistic framework that closely follows human behaviour, formally and explicitly dealing simultaneously with perceptual uncertainty, multisensory fusion and the perception–action loop.

Overview of the hierarchical Bayesian framework for active robotic perception

In the BVM framework, cells of a partitioning grid on the BVM log-spherical space Y associated with the egocentric coordinate system $\{E\}$ are indexed through $C \in Y$, representing the subset of positions in Y corresponding to the "far corners" ($log_b\, \rho_{\max}$, $\theta_{\max}$, $\varphi_{\max}$) 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 with the BVM for cell C, assuming a constant velocity hypothesis, as depicted on Fig. 2. Sensor measurements (i.e., the result of visual and auditory processing) are denoted by Z—

Journal : Large 10339 | Dispatch : 2-7-2012 | Pages : 5
Article No. : 481 | LE | TYPESET
MS Code : | CP | DISK

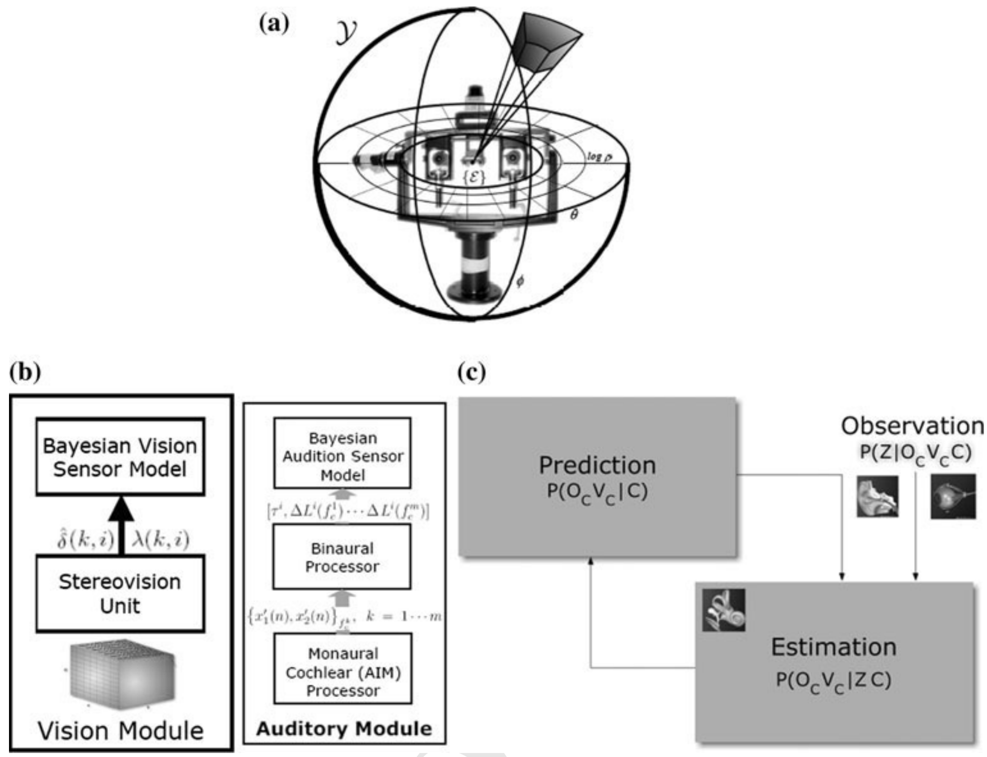

Fig. 2 Multisensory perception framework details (Ferreira et al. 2012). **a** The Bayesian volumetric map (BVM) referred to the egocentric coordinate frame of the robotic active perception system; **b** BVM sensor models; **c** BVM Bayesian occupancy filter

observations $P\ (Z\ |O_C\ V_C\ C)$ are given by the Bayesian sensor models of Fig. 2, which yield results already integrated within the log-spherical configuration.

The BVM framework is extensible in such a way that other properties characterised 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). 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.

Three decision models were proposed by Ferreira et al. (2012): π_A, which implements entropy-based active exploration based on the BVM; π_B, which uses entropy and saliency together for active perception; and finally, π_C, which adds a simple inhibition of return (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 Fig. 1. Each decision model will infer a probability distribution on the next point of fixation for the next desired gaze shift represented by a random variable $G_t \in Y$ at each time $t \in [1, t_{\max}]$. For more details, refer to (Ferreira et al. 2012).

The complete set of variables that set up the framework and its extensions is summarised in the following list (references to temporal properties 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 (OC, VC), 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 inhibition of return behaviour, ranging from no inhibition (0) to full inhibition (1).

Journal : Large 10339 | Dispatch : 2-7-2012 | Pages : 5
Article No. : 481 | LE | TYPESET
MS Code : | CP | DISK

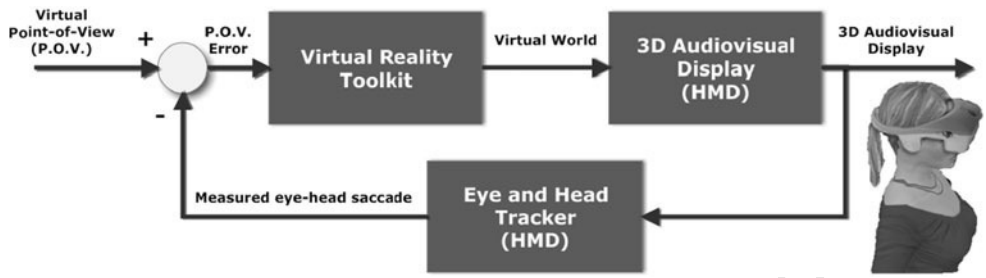

Fig. 3 Virtual point-of-view generator set-up that allows the updating of audiovisual stimuli presentation according to the monitored subjects' gaze direction

Learning automatic emergent behaviours for active multisensory perception

Any of the three decision models, π_A, π_B and π_C, results in an inference result similar to the following equation (which, in fact, corresponds to model π_B),

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

The automatic multisensory active perception behaviours emerge from the distributions $P(Q^{i,t}\ |G^t\ \pi_B)$, which are either beta distributions $B(\alpha_Q,\ \ \beta_Q)$ for $[G^t = C]$ expressing that, for a given point of fixation proposal for the next gaze shift, $Q_C^{i,t}$ is more likely near 1, or a uniform distribution on $Q_C^{i,t}$ for $[G^t \neq C]$. In this equation, the probability of a perceptually salient object occupying cell C, given by $Q_C^{i,t}$, is to be replaced by U_C^t or R_C^t, depending on which model besides π_B one is referring to, either π_A or π_C, respectively.

Therefore, the learning process in this context is defined as supervised learning through the maximum likelihood estimation (MLE) of the free parameters of the respective beta distributions. The training data to perform this learning is gathered from psychophysical experiments, in which human subjects using a head-mounted device are presented with realistic 3D, audiovisual, virtual-reality scenarios. The subjects' tracked head-eye gaze shifts control the virtual stereoscopic-binaural point of view and hence the progression of each stimulus movie—see Fig. 3—while audiovisual stimuli and corresponding fixation points are logged. This way, controlled conditions will be enforced by proposing both free-viewing and active search tasks to the subjects, thus enabling as systematic estimation of distribution parameters 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.

Discussion

At the time of writing this text, pilot experiments have already been conducted, validating the learning procedure and already displaying promising results.

In the following months, the final experiments will be conducted and a robotic demonstrator will be set up, using the learned attentional sets in implementing different tasks. This will further prove that the Bayesian hierarchical framework adequately follows 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.

Adaptivity—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).

Acknowledgments The authors would particularly like to thank, at the Institute of Biomedical Research in Light and Image of the University of Coimbra (IBILI/UC), Prof. Miguel Castelo-Branco, João Castelhano, Carlos Amaral and Marco Simões for their help with the psychophysical experiments.

Conflict of interest This supplement was not sponsored by outside commercial interests. It was funded entirely by ECONA, Via dei Marsi, 78, 00185 Roma, Italy.

Journal : Large 10339 | Dispatch : 2-7-2012 | Pages : 5
Article No. : 481 | LE | TYPESET
MS Code : | CP | DISK

References

Aloimonos J, Weiss I, Bandyopadhyay A (1987) Active vision. Int J Comput Vis 1:333–356

Bajcsy R (1985) Active perception vs passive perception. In: Third IEEE workshop on computer vision, Bellair, Michiganm, pp 55–59

Belardinelli A, Pirri F, Carbone A (2007) Bottom-up gaze shifts and fixations learning by imitation. IEEE Trans Syst Man Cybern B 37(2):256–271

Bessière P, Laugier C, Siegwart R (eds) (2008) Probabilistic reasoning and decision making in sensory-motor systems, volume 46 of Springer tracts in advanced robotics, Springer. ISBN: 978-3-540-79006-8

Breazeal C, Edsinger A, Fitzpatrick P, Scassellati B (2001) Active vision for sociable robots. IEEE Trans Syst Man Cybern A Syst Hum 31(5):443–453

Buswell GT (1935) How people look at pictures: a study of the psychology and perception in art. University Chicago Press, Chicago

Castelhano MS, Mack ML, Henderson JM (2009) Viewing task influences eye movement control during active scene perception. J Vis 9:1–15

Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci 3(3):201–215

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

Ferreira JF (2011) Bayesian cognitive models for 3D structure and motion multimodal perception. PhD thesis, Faculty of Sciences and Technology of the University of Coimbra (FCTUC)

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

Ferreira JF, Castelo-Branco M, Dias J (2012) A hierarchical Bayesian framework for multimodal active perception. Adapt Behav 20(3):172–190 Published online ahead of print, March 1

Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259

Mills M, Hollingworth A, Dodd MD (2011) Examining the influence of task set on eye movements and fixations. J Vis 11:1–15

Journal : Large 10339 Dispatch : 2-7-2012 Pages : 5
Article No. : 481 ☐ LE ☐ TYPESET
MS Code : ☑ CP ☑ DISK