A Bayesian Hierarchical Framework for Multimodal Active Perception

João Filipe Ferreira and Jorge Dias

Institute of Systems and Robotics,
FCT-University of Coimbra
Coimbra, Portugal
{jfilipe,jorge}@isr.uc.pt
http://paloma.isr.uc.pt

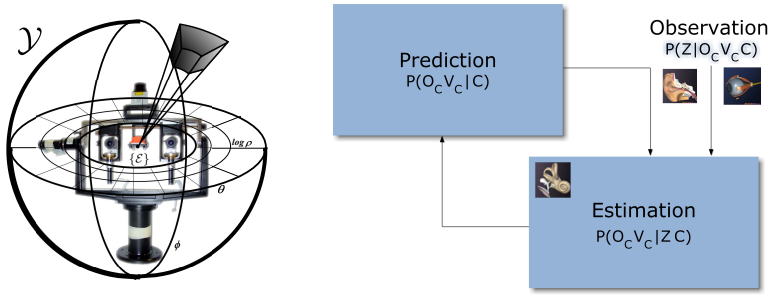
Abstract. We will present a Bayesian hierarchical framework for multimodal active perception, devised to be *emergent*, *scalable* and *adaptive*, together with some representative experimental results. This framework, while not strictly neuromimetic, finds its roots in the role of the dorsal perceptual 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.

Keywords: Bayesian approach, multimodal perception, emergence, adaptivity, scalability.

1 Introduction

Active perception has been an object of study in robotics for decades now, specially active vision, which was first introduced by [2] and later explored by [1]. Many perceptual tasks tend to be simpler if the observer is active and controls its sensors [1]. 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 [1].

We will present a complex artificial active perception system that follows human-like bottom-up driven behaviours using vision, audition and vestibular sensing. More specifically, the conceptual tool of Bayesian Programming [3] 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 [9,10] and automatic orientation based on sensory saliency [12]. A real-time implementation of all the processes of the framework has been developed, capitalising on the potential for parallel computing of most of its algorithms, as an extension of what was presented in [8]. An overview of the framework and its models will be summarised in this text, and representative results will be presented. In the process, we will demonstrate the following properties which are intrinsic to the framework: *emergence*, *scalability* and *adaptivity*.



(a) (b)

Fig. 1. Multimodal perception framework details. (a) The Bayesian Volumetric Map (BVM) and the Integrated Multimodal Perception Experimental Platform (IMPEP); (b) BVM Bayesian Occupancy Filter.

2 Active Multimodal Perception — A Hierarchy of Bayesian Models

To achieve our goal of developing Bayesian models for visuoauditory-driven saccade generation, a *representation model* and *decision models* for gaze shift generation, similar to what was proposed by Colas *et al.* [4], have been developed and are summarised in the following text.

A spatial representation framework for multimodal perception of 3D structure and motion, the Bayesian Volumetric Map (BVM), was presented in [7], characterised by an *egocentric*, *log-spherical spatial configuration* to which the Bayesian Occupancy Filter (BOF), as formalised by Tay *et al.* [13], has been adapted (see Fig. 1).

In this model, cells of a partitioning grid on the BVM log-spherical space $\mathcal{Y}$ are indexed through $C \in \mathcal{C} \subset \mathcal{Y}$, where $\mathcal{C}$ represents 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 [6]), 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 Fig. 1. 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 Bayesian sensor models, which yield results already integrated within the log-spherical configuration, as presented in [7].

The BVM 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, by augmenting the hierarchy of operators through Bayesian subprogramming [3,11]. This ensures that the framework is *scalable*.

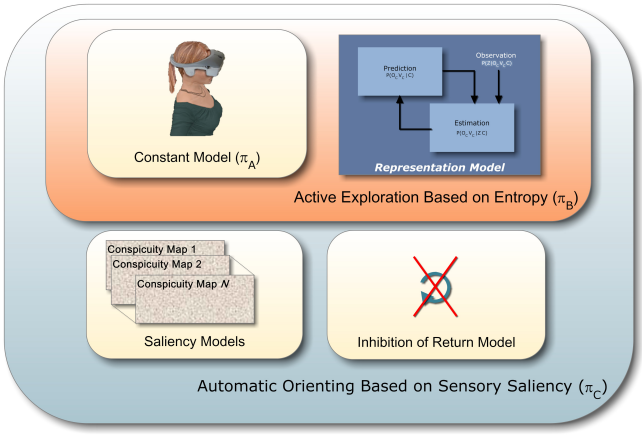


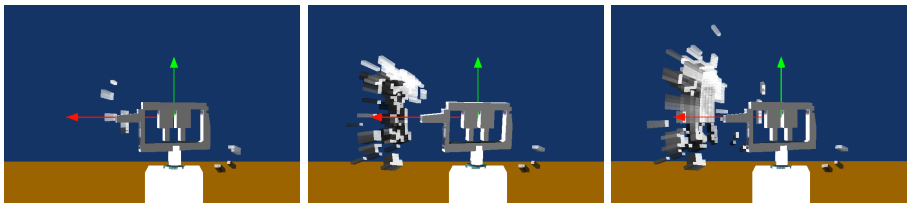
Fig. 2. Active perception model hierarchy.

Therefore, we will specify two decision models[1]: one that implements entropy-based active exploration using the representation model (π_B), and one that uses entropy and saliency together for active perception (π_C). In other words, each model π_k incorporates its predecessor π_{k-1} through Bayesian fusion, therefore constituting a model hierarchy — see Fig. 2. The first model we propose uses the knowledge from the representation layer to determine gaze shift fixation points. More precisely, it tends to look towards locations of high entropy/uncertainty. Its likelihood will be 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 as described in [9,10]. The second model is given by the product between the prior on gaze shifts due to entropy-based active exploration, the Inhibition of Return (IoR) model [12], and each distribution on sensory-salient BVM cells. This expression shows that the model is attracted towards both salient cells *and* locations of high uncertainty, while avoiding the fixation site computed on the previous time step through the IoR process — the combination of these strategies to produce a coherent behaviour ensures that the framework is *emergent*. The parameters used for each distribution may be introduced directly by the programmer (like a genetic imprint) or they may be manipulated "on the fly", which in turn would allow for goal-dependent behaviour implementation (i.e. top-down influences), and therefore ensure that the framework is *adaptive*.

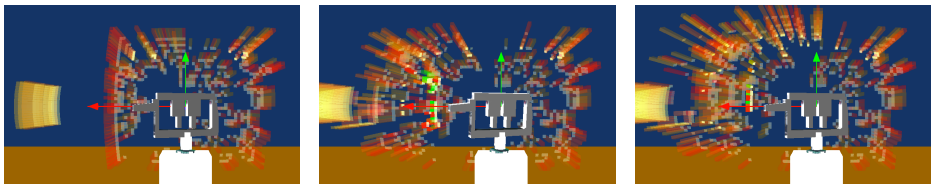
3 Results

Experimental results showcasing some of the capabilities of the system are presented in Fig. 3.

[1] Constant model π_A is for Bayesian learning purposes of system parameters, and is beyond the scope of this text — it is considered herewith as a uniform prior.



(a) BVM results corresponding to a scenario composed of one male speaker calling out at approximately 30° azimuth relatively to the Z axis, which defines the frontal heading respective to the IMPEP "neck". The reconstruction of the speaker can be clearly seen on the left of each representation of the BVM.



(b) Relevant values for entropy-based variable U_C corresponding to each of the time-instants in (a). Represented values range from .5 to 1, depicted using a smoothly gradated red-to-green colour-code (red corresponds to lower values, green corresponds to higher values). Chronologically ordered interpretation of these results goes as follows: at first, relevant cells have their relative importance for sensory exploration scattered throughout the visible area, and there is a separate light yellow region on the left corresponding to an auditory object (i.e the speaker) that becomes the focus of interest; then, at the boundaries of the speaker's silhouette, bright green cells show high relevance of this area for exploration, which then becomes the next focus of interest; finally, after a few cycles of BVM processing, uncertainty lowers, which clearly shows as the number of green cells diminishes.

Fig. 3. Results corresponding, from left to right, to time-instants in which gaze shifts were generated, 17.080 s, 26.664 s and 36.411 s, respectively, exemplifying the use of the entropy-based variable U_C to elicit gaze shifts, in order to scan the surrounding environment. A scene consisting of a male speaker talking in a cluttered lab is observed by the IMPEP active perception system and processed online by the Bayesian framework. An oriented 3D avatar of the IMPEP perception system depicted in each map 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, $\theta \in [-180^{\circ}, 180^{\circ}]$, with $\Delta\theta = 1^{\circ}$, and $\phi \in [-90^{\circ}, 90^{\circ}]$, with $\Delta\phi = 1^{\circ}$, 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 [5]).

Acknowledgments This publication has been supported by the European Commission within the *Seventh Framework Programme FP7, as part of theme 2: Cognitive Systems, Interaction, Robotics, under grant agreement 231640*. The contents of this text reflect only the author's views. The European Community is not liable for any use that may be made of the information contained herein.

References

1. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active Vision. International Journal of Computer Vision 1, 333–356 (1987) 1
2. Bajcsy, R.: Active perception vs passive perception. In: Third IEEE Workshop on Computer Vision. pp. 55–59. Bellair, Michigan (1985) 1
3. Bessière, P., Laugier, C., Siegwart, R. (eds.): Probabilistic Reasoning and Decision Making in Sensory-Motor Systems, Springer Tracts in Advanced Robotics, vol. 46. Springer (2008), iSBN: 978-3-540-79006-8 1, 2
4. Colas, F., Flacher, F., Tanner, T., Bessière, P., Girard, B.: Bayesian models of eye movement selection with retinotopic maps. Biological Cybernetics 100, 203–214 (2009) 2
5. Cutting, J.E., Vishton, P.M.: Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In: Epstein, W., Rogers, S. (eds.) Handbook of perception and cognition, vol. 5; Perception of space and motion. Academic Press (1995) 4
6. Elfes, A.: Using occupancy grids for mobile robot perception and navigation. IEEE Computer 22(6), 46–57 (1989) 2
7. Ferreira, J.F., Bessière, P., Mekhnacha, K., Lobo, J., Dias, J., Laugier, C.: 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 (April 2008) 2
8. Ferreira, J.F., Lobo, J., Dias, J.: Bayesian Real-Time Perception Algorithms on GPU — Real-Time Implementation of Bayesian Models for Multimodal Perception Using CUDA. Journal of Real-Time Image Processing Special Issue (February 26 2010), springer Berlin/Heidelberg, published online (ISSN: 1861-8219) 1
9. Ferreira, J.F., Pinho, C., Dias, J.: Active Exploration Using Bayesian Models for Multimodal Perception. In: Campilho, A., Kamel, M. (eds.) Image Analysis and Recognition, Lecture Notes in Computer Science series (Springer LNCS), International Conference ICIAR 2008. pp. 369–378 (June 25–27 2008) 1, 3
10. Ferreira, J.F., Prado, J., Lobo, J., Dias, J.: Multimodal Active Exploration Using A Bayesian Approach. In: IASTED International Conference in Robotics and Applications. pp. 319–326. Cambridge MA, USA (November 2–4 2009) 1, 3
11. Lebeltel, O.: Programmation Bayésienne des Robots. Ph.D. thesis, Institut National Polytechnique de Grenoble, Grenoble, France (September 1999) 2
12. Niebur, E., Itti, L., Koch, C.: Modeling the "where´´ visual pathway. In: Sejnowski, T.J. (ed.) 2nd Joint Symposium on Neural Computation, Caltech-UCSD. vol. 5, pp. 26–35. Institute for Neural Computation, La Jolla (1995) 1, 3
13. Tay, C., Mekhnacha, K., Chen, C., Yguel, M., Laugier, C.: An efficient formulation of the bayesian occupation filter for target tracking in dynamic environments (2007), *International Journal of Autonomous Vehicles* 2