Active Exploration Using Bayesian Models for Multimodal Perception

João Filipe Ferreira, Cátia Pinho, and Jorge Dias*

ISR — Institute of Systems and Robotics, FCT-University of Coimbra
Coimbra, Portugal

Abstract. In this text we will present a novel solution for active perception built upon a probabilistic framework for multimodal perception of 3D structure and motion — the Bayesian Volumetric Map (BVM). This solution applies the notion of entropy to promote gaze control for active exploration of areas of high uncertainty on the BVM so as to dynamically build a spatial map of the environment storing the largest amount of information possible. Moreover, entropy-based exploration is shown to be an efficient behavioural strategy for active multimodal perception.

1 Introduction

Perception has been regarded as a computational process of unconscious, probabilistic inference. Aided by developments in statistics and artificial intelligence, researchers have begun to apply the concepts of probability theory rigorously to problems in biological perception and action. One striking observation from this work is the myriad ways in which human observers behave as near-optimal Bayesian observers, which has fundamental implications for neuroscience, particularly in how we conceive of neural computations and the nature of neural representations of perceptual variables [1].

Consider the following scenario — an observer is presented with a non-static 3D scene containing several moving entities, probably generating some kind of sound: how does this observer perceive the 3D structure of all entities in the scene and the 3D trajectory and velocity of moving objects, given the ambiguities and conflicts inherent to the perceptual process? Given these considerations, the research presented on this text regards a Bayesian framework for artificial multimodal perception models.

In this text we will present a novel solution for active perception built upon a probabilistic framework for multimodal perception of 3D structure and motion — the Bayesian Volumetric Map, a metric, egocentric spatial memory. This solution

* This publication has been supported by EC-contract number *FP6-IST-027140, Action line: Cognitive Systems*. 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.

A. Campilho and M. Kamel (Eds.): ICIAR 2008, LNCS 5112, pp. 369–378, 2008.

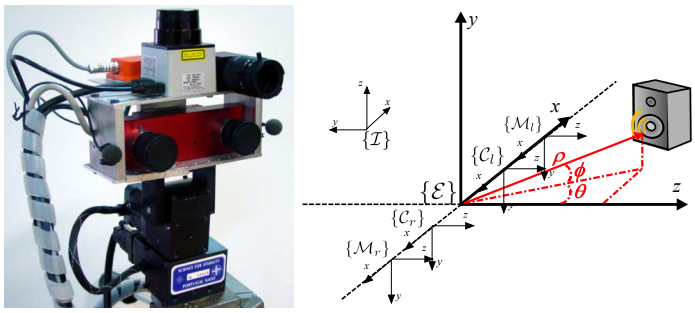

Fig. 1. View of the first version of the Integrated Multimodal Perception Experimental Platform (IMPEP), on the left. On the right, the IMPEP perceptual geometry is shown: $\{\mathcal{E}\}$ is the main reference frame for the IMPEP robotic head, representing the egocentric coordinate system;$\{\mathcal{C}_{l,r}\}$ are the stereovision (respectively left and right) camera referentials; $\{\mathcal{M}_{l,r}\}$ are the binaural system (respectively left and right) microphone referentials; and finally $\{\mathcal{I}\}$ is the inertial measuring unit's coordinate system.

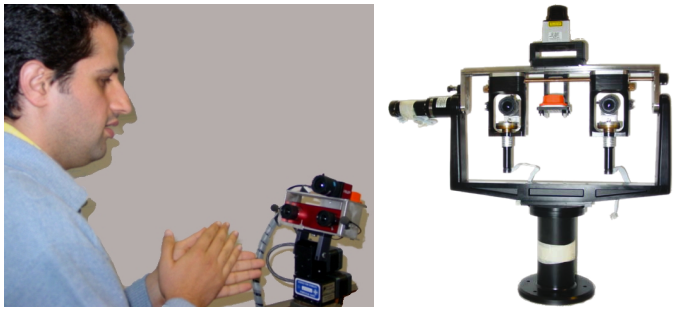

Fig. 2. On the left: typical application context of the IMPEP active perception system. On the right, a perspective of the current version of the IMPEP active perception head, which adds vergence capabilities to the stereovision system besides improved motor control and conditioning.

applies the notion of entropy to promote gaze control for active exploration of areas of high uncertainty on the BVM so as to dynamically build a spatial map of the environment storing the largest amount of information possible.

To support our research work, an artificial multimodal perception system (IMPEP — Integrated Multimodal Perception Experimental Platform) has been constructed at the ISR/FCT-UC consisting of a stereovision, binaural and inertial measuring unit (IMU) setup mounted on a motorised head, with gaze control capabilities for image stabilisation and perceptual attention purposes — see Figs. 1 and 2. This solution will enable the implementation of an active perception system with great potential in applications as diverse as social robots or even robotic navigation (Fig. 2).

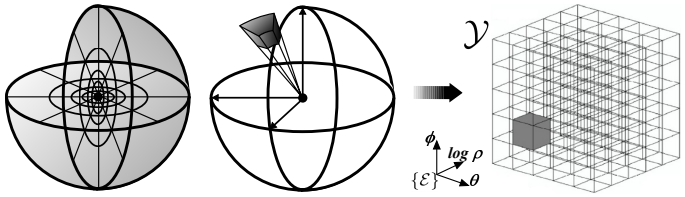

Fig. 3. Egocentric, log-spherical configuration of the Bayesian Volumetric Maps

2 Bayesian Volumetric Maps for Multimodal Perception

2.1 Volumetric Map Spatial Configuration Definition

Metric maps are very intuitive, yield a rigorous model of the environment and help to register measurements taken from different locations. Grid-based maps are the most popular metric maps in mobile robotics applications. One of the most popular grid-based maps is the *occupancy grid*, which is a discretised random field where the probability of occupancy of each cell is kept, and the probability values of occupancy of all cells *are considered independent between each other* [2]. The absence of an object based representation permits the ease of fusing low level descriptive sensory information onto the grids without necessarily implicating data association.

We have developed a *log-spherical* coordinate system grid (see Fig. 3) that promotes an egocentric trait and yields more precision for objects closer to the observer, which seems to agree with biological perception.

This spatial configuration is primarily defined by its range of azimuth and elevation angles, and by its maximum reach in distance ρ_{Max}, which in turn determines its log-distance base through $b = a^{\frac{\log_a(\rho_{\mathrm{Max}} - \rho_{\mathrm{Min}})}{N}}, \forall a \in \mathbb{R}$, where ρ_{Min} defines the *egocentric gap*, for a given number of partitions N, chosen according to application requirements. This space is therefore effectively defined by

$$\mathcal{Y} \equiv]\log_b \rho_{\mathrm{Min}}; \log_b \rho_{\mathrm{Max}}] \times]\theta_{\mathrm{Min}}; \theta_{\mathrm{Max}}] \times]\phi_{\mathrm{Min}}; \phi_{\mathrm{Max}}] \tag{1}$$

In practice, this grid is parametrised so as to cover the full angular range for azimuth and elevation.

Each cell of the grid is defined by two limiting log-distances, $\log_b \rho_{\mathrm{min}}$ and $\log_b \rho_{\mathrm{max}}$, two limiting azimuth angles, θ_{min} and θ_{max}, and two limiting elevation angles, ϕ_{min} and ϕ_{max}, through:

$$\mathcal{Y} \supset \mathcal{C} \equiv]\log_b \rho_{\mathrm{min}}; \log_b \rho_{\mathrm{max}}] \times]\theta_{\mathrm{min}}; \theta_{\mathrm{max}}] \times]\phi_{\mathrm{min}}; \phi_{\mathrm{max}}] \tag{2}$$

where constant values for log-distance base b, and angular ranges $\Delta\theta = \theta_{\mathrm{max}} - \theta_{\mathrm{min}}$ and $\Delta\phi = \phi_{\mathrm{max}} - \phi_{\mathrm{min}}$, chosen according to application resolution requirements, ensure grid regularity. Finally, each cell is formally *indexed* by the coordinates of its *far corner*, defined as $C = (\log_b \rho_{\mathrm{max}}, \theta_{\mathrm{max}}, \phi_{\mathrm{max}})$.

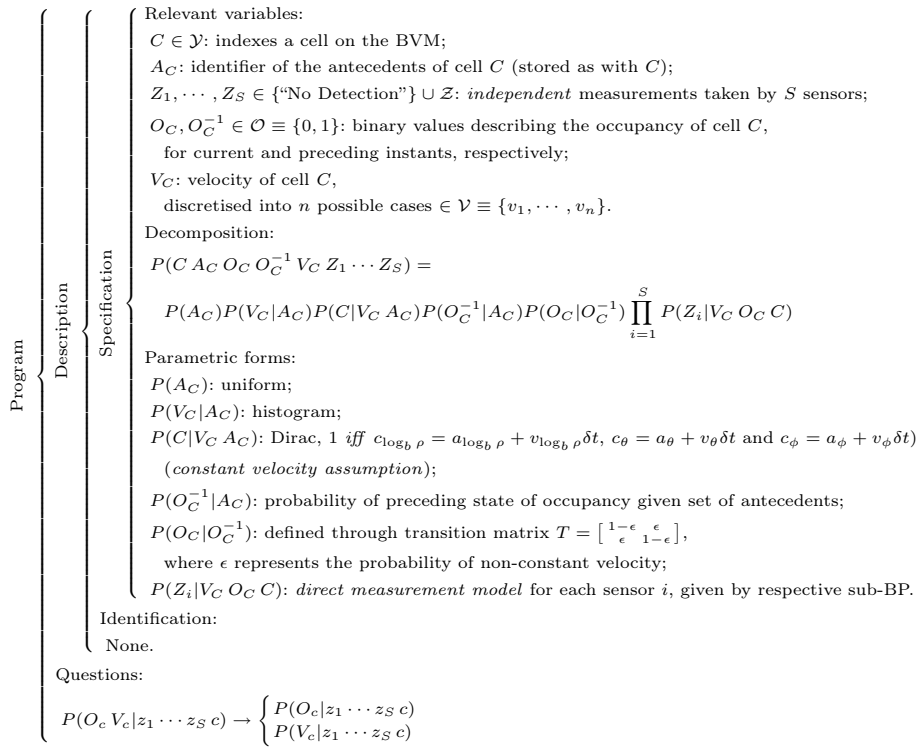

Fig. 4. Bayesian Program for the estimation of Bayesian Volumetric Map current cell state

2.2 Using Bayesian Filtering for Visuoauditory Integration

We have developed a novel probabilistic, volumetric occupancy grid framework called the *Bayesian Volumetric Map* (BVM), which is defined in the Bayesian Program presented in Fig. 4, a formalism created by Lebeltel [3] to supersede, restate and compare numerous classical probabilistic models such as Bayesian Networks (BN), Dynamic Bayesian Networks (DBN), Bayesian Filters, Hidden Markov Models (HMM), Kalman Filters, Particle Filters, Mixture Models, or Maximum Entropy Models. The BVM is based on the solution presented by Tay *et al.* [4] called the Bayesian Occupancy Filter (BOF), adapted so as to conform to the BVM egocentric, three-dimensional and log-spherical nature.

The estimation of the joint state of occupancy and velocity of a cell is answered through Bayesian inference on the decomposition equation given in Fig. 4. This inference effectively leads to the Bayesian filtering formulation as used in the BOF grids — see Fig. 5. In this context, prediction propagates cell occupancy probabilities for each velocity and cell in the grid — $P(O_C\, V_C|C)$. During estimation, $P(O_C\, V_C|C)$ is updated by taking into account the observations yielded by the sensors $\prod_{i=1}^{S} P(Z_i|V_C\, O_C\, C)$ to obtain the final state estimate

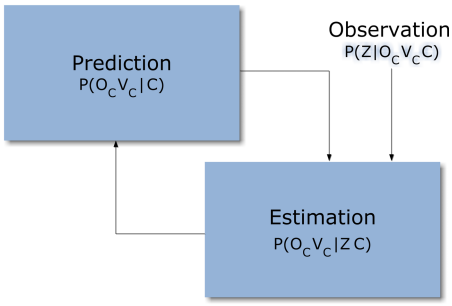

Fig. 5. Bayesian filtering for the estimation of occupancy and local motion distributions in the BVM. The schematic considers only a single measurement for simpler reading, with no loss of generality.

$P(O_C\, V_C | Z_1 \cdots Z_S\, C)$. The result from the Bayesian filter estimation will then be used for the prediction step in the next iteration.

2.3 Using the BVM for Sensory Combination of Vision and Audition with Vestibular Sensing

Consider the simplest case, where the sensors may only rotate around the egocentric origin and the whole perceptual system is not allowed to perform any translation. In this case, the vestibular sensor models 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.

To maintain a head-centred coordinate system for the BVM, which would obviously shift in accordance to head turns, instead of rotating the whole map, the most effective solution 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* $C = (\log_b \rho_{max}, \theta_{max} - \theta_{inertial}, \phi_{max} - \phi_{inertial}) \in \mathcal{C} \subset \mathcal{Y}$.

This process obviously relies on the assumption that inertial precision on angular measurements is greater than the chosen resolution parameters for the BVM.

2.4 Sensor Models

Our motivations suggest for the vision sensor model a tentative data structure analogous to neuronal population activity patterns to represent uncertainty in the form of probability distributions — a spatially organised 2D grid has each cell associated to a population code simulation, a set of probability values of a neuronal population encoding a probability distribution [5]. The stereovision algorithm used for visual depth sensing is an adaptation of the fast and simple coherence detection approach by Henkel [6], yielding an estimated disparity map $\hat{\delta}(k, i)$ and a corresponding confidence map $\lambda(k, i)$. For visual perception of occupancy, this stereovision sensor described can be decomposed into simpler

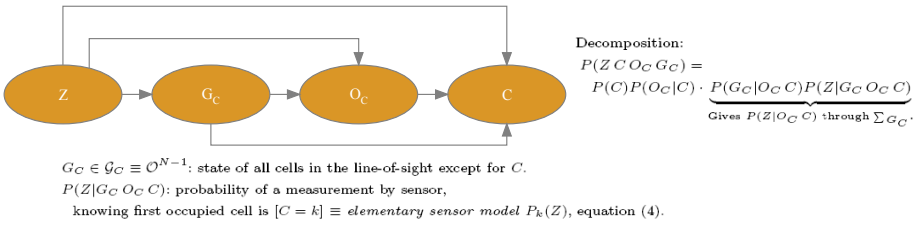

Fig. 6. Bayes network corresponding to the Bayesian Program for the vision sensor model. Variables and distributions other than the ones already defined on the Bayesian Program of Fig. 4 are presented below the diagram.

linear (1D) depth $\rho(k,i)$ measuring sensors per projection line/pixel (k,i), each oriented in space with spherical angles $(\theta(k,i), \phi(k,i))$.

This algorithm is then easily converted from its deterministic nature into a probabilistic implementation simulating the population code-type data structure. This results in probability distributions on sensor measurements made available as likelihood functions taken from sensor readings — *soft evidence*, or "Jeffrey's evidence" in reference to Jeffrey's rule [7]; the relation between vision sensor measurements Z and the corresponding readings δ and λ is thus described by the calibrated expected value $\hat{\rho}(\hat{\delta})$ and standard deviation $\sigma_\rho(\lambda)$ for each sensor, defined later on.

We have decided to model these sensors in terms of their contribution to the estimation of cell occupancy in a similar fashion to the solution proposed by Yguel *et al.* [8] — see the Bayesian Program presented on Fig. 6. The answer to the Bayesian Program question in order to determine the sensor model $P(Z|O_C\,C)$ for vision, which is in fact related to the decomposition of interest $P(O_C\,Z\,C) = P(C)P(O_C|C)P(Z|O_C\,C)$, is answered through Bayesian inference on the decomposition equation; the inference process will dilute the effect of the unknown probability distribution $P(G_C|O_C\,C)$ through marginalisation over all possible states of G_C. In other words, the resulting *direct* model for vision sensors is based solely on knowing which is the first occupied cell on the line-of-sight and its relative position to a given cell of interest C.

Given the first occupied cell $[C = k]$ on the line-of-sight, the likelihood functions yielded by the population code data structure become

$$P_k(Z) = L_k(Z, \mu_\rho(k), \sigma_\rho(k)), \begin{cases} \mu_\rho(k) & = \hat{\rho}(\hat{\delta}) \\ \sigma_\rho(k) & = \frac{1}{\lambda}\sigma_{min} \end{cases} \tag{3}$$

with σ_{min} and $\hat{\rho}(\hat{\delta})$ taken from calibration, the former as the estimate of the smallest error in depth yielded by the stereovision system and the latter from the intrinsic camera geometry. The likelihood function *constitutes, in fact, the elementary sensor model* as defined above for each vision sensor.

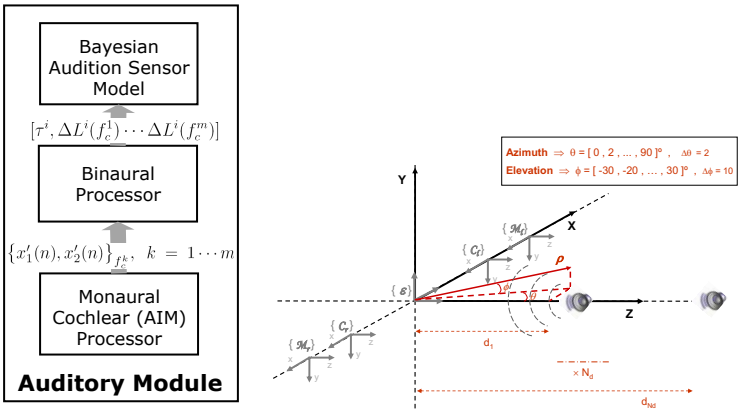

Fig. 7. On the left: the IMPEP Bayesian binaural system. On the right: schematic of a typical auditory sensor space configuration defined during calibration.

We have adapted Yguel *et al.*'s Gaussian elementary sensor model so as to additionally perform the transformation to distance log-space, as follows

$$P_k([Z = z]) = \begin{cases} \int_{]-\infty;1]} \mathcal{N}(\mu(k-0.5), \sigma(\sigma_\rho))(u)du, & z \in [0;1] \\ \int_{\lfloor z \rfloor}^{\lfloor z \rfloor + 1} \mathcal{N}(\mu(k-0.5), \sigma(\sigma_\rho))(u)du, & z \in]1;N] \\ \int_{]N;+\infty]} \mathcal{N}(\mu(k-0.5), \sigma(\sigma_\rho))(u)du, & z = \text{“No Detection”} \end{cases} \tag{4}$$

where $\mu(\bullet)$ and $\sigma(\bullet)$ are the operators that perform the required spatial coordinate transformations, and $k = \lceil \mu_\rho \rceil$ is assumed to be the log-space index of the only occupied cell in the line-of-sight, which represents the coordinate interval $]k-1;k]$.

As for the audition sensor model, it is built upon a binaural sensor system composed of three distinct and consecutive processors (Fig. 7): 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 [9]; 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 localisation unit*, which applies a Bayesian sensor model so as to perform localisation of sound-sources in 3D space. A full description together with preliminary results have been presented in [10].

To process the inertial data, we adapted the Bayesian model proposed by Laurens and Droulez [11] for the human vestibular system. The aim is to provide an estimate for the current angular position and angular velocity of the system, that mimics human vestibular perception.

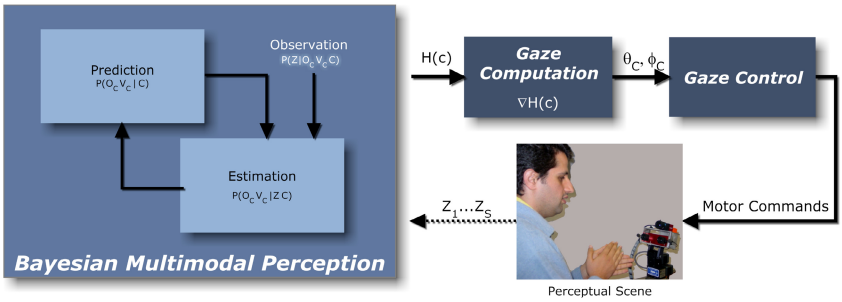

Fig. 8. Active multimodal perception using entropy-based exploration. Only the Bayesian models for multimodal perception and the entropy-based exploration algorithm implemented by the gaze computation module are described herewith; the gaze control module is beyond the scope of this text.

3 Active Multimodal Perception Using Entropy-Based Exploration

Active perception has been an object of study in robotics for decades now, specially active vision, which was first introduced by Bajcsy *et al.* [12] and later explored by Aloimonos *et al.* [13]. Many perceptual tasks tend to be simpler if the observer is active and controls its sensors [13]. 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 [13]. Moreover, the research of behavioural strategies to implement active perception as efficiently as possible is a very important research topic.

The availability of a probabilistic framework to implement spatial mapping of the environment substantiated by the BVM allows the use of the concept of *information entropy*, which can be used to promote an exploratory behaviour of areas of the environment corresponding to cells on the volumetric map associated to high uncertainty, an idea recently explored by Rocha *et al.* [14,15].

Information in the BVM is stored as the *probability of each cell being in a certain state*, defined in the BP of Fig. 5 as $P(V_c\, O_c | z\, c)$. The state of each cell thus belongs to the state-space $\mathcal{O} \times \mathcal{V}$. The *joint entropy* of the random variables V_C and O_C that compose the state of each BVM cell $[C = c]$ is defined as follows:

$$H(c) \equiv H(V_c, O_c) = -\sum_{\substack{o_c \in \mathcal{O} \\ v_c \in \mathcal{V}}} P(v_c\, o_c | z\, c) \log P(v_c\, o_c | z\, c) \tag{5}$$

The joint entropy value $H(c)$ is a sample of a continuous joint entropy field $H : \mathcal{Y} \rightarrow \mathbb{R}$, taken at log-spherical positions $[C = c] \in \mathcal{C} \subset \mathcal{Y}$. Let $c_{\alpha-}$ denote the contiguous cell to C along the negative direction of the generic log-spherical

axis α, and consider the edge of cells to be of unit length in log-spherical space, without any loss of generality. A reasonable first order approximation to the joint entropy gradient at $[C = c]$ would be

$$\overrightarrow{\nabla} H(c) \approx [H(c) - H(c_{\rho-}), H(c) - H(c_{\theta-}), H(c) - H(c_{\phi-})]^T \tag{6}$$

with magnitude $\|\overrightarrow{\nabla} H(c)\|$.

A great advantage of the BVM over Cartesian implementations of occupancy maps such as the one presented on [14,15] is the fact that the log-spherical configuration avoids the need for time-consuming ray-casting techniques when computing a gaze direction for active exploration, since the log-spherical space is already defined based on directions (θ, ϕ). Hence, the active exploration algorithm is simplified to the completion of the following steps:

1. Find the last non-occluded, close-to-empty (i.e. $P([O_C = 1]|[C = c]) < .5$) cell for the whole span of directions $(\theta_{\max}, \phi_{\max})$ in the BVM — these are considered to be the so-called *frontier cells* as defined on [14,15]; the set of all frontier cells will be denoted here as $\mathcal{F} \subset \mathcal{C}$.
2. Compute the joint entropy gradient for each of the frontier cells and select $c_s = \arg\max_{c \in \mathcal{F}} \left[(1 - P([O_C = 1]|[C = c])) \|\overrightarrow{\nabla} H(c)\| \right]$ as the best candidate cell to direct gaze to. In case there is more than one global maximum, choose the cell corresponding to the direction closest to the current heading (i.e. $(\theta_{\max}, \phi_{\max}) = (0, 0)$, so as to ensure minimum gaze shift rotation effort.
3. Compute gaze direction as being (θ_C, ϕ_C), where θ_C and ϕ_C are the angles that bisect cell $[C = c_s]$ (i.e. which pass through the geometric centre of cell c_s in Cartesian space).

The full BVM entropy-based active perception system is described by the block diagram presented in Fig. 8.

4 Conclusions

In this text we have presented a novel solution for active perception built upon a probabilistic framework for multimodal perception of 3D structure and motion — the Bayesian Volumetric Map (BVM), a metric, egocentric spatial memory. This solution applies the notion of entropy to promote gaze control for active exploration of areas of high uncertainty on the BVM so as to dynamically build a spatial map of the environment storing the largest amount of information possible. Moreover, entropy-based exploration was shown to be an efficient behavioural strategy for active multimodal perception.

Further details on the calibration and implementation of these models on the Integrated Multimodal Perception Experimental Platform can be found at `http://paloma.isr.uc.pt/~jfilipe/BayesianMultimodalPerception`.

References

1. Knill, D.C., Pouget, A.: The Bayesian brain: the role of uncertainty in neural coding and computation. TRENDS in Neurosciences 27(12), 712–719 (2004)
2. Elfes, A.: Using occupancy grids for mobile robot perception and navigation. IEEE Computer 22(6), 46–57 (1989)
3. Lebeltel, O.: Programmation Bayésienne des Robots. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France (September 1999)
4. Tay, C., Mekhnacha, K., Chen, C., Yguel, M., Laugier, C.: An efficient formulation of the bayesian occupation filter for target tracking in dynamic environments. International Journal of Autonomous Vehicles (2007)
5. Pouget, A., Dayan, P., Zemel, R.: Information processing with population codes. Nature Reviews Neuroscience 1, 125–132 (2000)
6. Henkel, R.: A Simple and Fast Neural Network Approach to Stereovision. In: Jordan, M., Kearns, M., Solla, S. (eds.) Proceedings of the Conference on Neural Information Processing Systems — NIPS 1997, pp. 808–814. MIT Press, Cambridge (1998)
7. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Revised second printing edn. Morgan Kaufmann Publishers, Inc. (Elsevier) (1988)
8. Yguel, M., Aycard, O., Laugier, C.: Efficient GPU-based Construction of Occupancy Grids Using several Laser Range-finders. International Journal of Autonomous Vehicles (2007)
9. Patterson, R.D., Allerhand, M.H., Giguère, C.: Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. J. AcoustS. Soc. Am, 1890–1894 (1995)
10. Pinho, C., Ferreira, J.F., Bessière, P., Dias, J.: A Bayesian Binaural System for 3D Sound-Source Localisation. In: International Conference on Cognitive Systems (CogSys 2008), University of Karlsruhe, Karlsruhe, Germany (April 2008)
11. Laurens, J., Droulez, J.: Bayesian processing of vestibular information. Biological Cybernetics (December 2006) (Published online: 5th December 2006)
12. Bajcsy, R.: Active perception vs passive perception. In: Third IEEE Workshop on Computer Vision, Bellair, Michigan, pp. 55–59 (1985)
13. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active Vision. International Journal of Computer Vision 1, 333–356 (1987)
14. Rocha, R., Dias, J., Carvalho, A.: Cooperative Multi-Robot Systems: a study of Vision-based 3-D Mapping using Information Theory. Robotics and Autonomous Systems 53(3–4), 282–311 (2005)
15. Rocha, R., Dias, J., Carvalho, A.: Exploring information theory for vision-based volumetric mapping. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005), Edmonton, Canada, August 2005, pp. 2409–2414 (2005)