A Bayesian Framework for Active Artificial Perception

Joao Ferreira, Jorge Lobo, Pierre Bessiere, M. Castelo-Branco, Jorge Dias

A Bayesian Framework for Active Artificial Perception

João Filipe Ferreira, *Member, IEEE,* Jorge Lobo, *Member, IEEE,* Pierre Bessière, Miguel Castelo-Branco, and Jorge Dias, *Senior Member, IEEE*

***Abstract*—In this text, we present a Bayesian framework for active multimodal perception of 3D structure and motion. The design of this framework finds its inspiration in the role of the dorsal perceptual pathway of the human brain. Its composing models build upon a common egocentric spatial configuration that is naturally fitting for the integration of readings from multiple sensors using a Bayesian approach. In the process, we will contribute with efficient and robust probabilistic solutions for cyclopean geometry-based stereovision and auditory perception based only on binaural cues, modelled using a consistent formalisation that allows their hierarchical use as building blocks for the multimodal sensor fusion framework. We will explicitly or implicitly address the most important challenges of sensor fusion using this framework, for vision, audition and vestibular sensing. Moreover, interaction and navigation requires maximal awareness of spatial surroundings, which in turn is obtained through active attentional and behavioural exploration of the environment. The computational models described in this text will support the construction of a simultaneously flexible and powerful robotic implementation of multimodal active perception to be used in real-world applications, such as human-machine interaction or mobile robot navigation.**

***Index Terms*—Sensing and Perception, Computer Vision, Sensor Fusion, Biologically-Inspired Robots, Multisensory Exploration, Active Perception, Multimodal Perception, Bayesian Programming.**

I. Introduction

Humans and robots alike have to deal with the unavoidable reality of sensory uncertainty. Consider the following scenario — a static or moving observer is presented with a non-static 3D scene containing several static and moving entities, probably generating some kind of sound: how does this observer perceive the 3D location, motion trajectory and velocity of all entities in the scene, while taking into account the ambiguities and conflicts inherent to the perceptual process?

Within the human brain, both dorsal and ventral visual systems process information about spatial location, but in very different ways: allocentric spatial information about how objects are laid out in the scene is computed by ventral stream mechanisms, while precise egocentric spatial information about the location of each object in a body-centred frame of reference is computed by the dorsal stream mechanisms, and also the phylogenetically preceding *superior colliculus* (SC), both of which mediate the perceptual control of action [1]. On the other hand, several authors argue that recent findings strongly suggest that the brain codes complex patterns of sensory uncertainty in its internal representations and computations [2, 3].

Finally, direction and distance in egocentric representations are believed to be separately specified by the brain [4, 5]. Considering distance in particular, just-discriminable depth thresholds have been usually plotted as a function of the log of distance from the observer, with analogy to contrast sensitivity functions based on Weber's fraction [6].

These findings inspired the construction of a probabilistic framework that allows fast processing of multisensory-based inputs to build a perceptual map of space so as to promote immediate action on the environment (as in the dorsal stream and superior colliculus), effectively postponing data association such as object segmentation and recognition to higher-level stages of processing (as in the ventral stream) — this would be analogous to a tennis player being required to hit a ball regardless of perception of its texture properties. Our framework bears a spherical spatial configuration (i.e. encoding 3D distance and direction), and constitutes a short-term perceptual memory performing efficient, lossless compression through log-partitioning of depth.

J. F. Ferreira (jfilipe@isr.uc.pt), J. Lobo and J. Dias are with the Institute of Systems and Robotics and the Faculty of Science and Technology, University of Coimbra, Coimbra, Portugal

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

P. Bessière is with the CNRS-Collège de France, Paris, France

M. Castelo-Branco is with the Biomedical Institute of Research on Light and Image, Faculty of Medicine, University of Coimbra, Coimbra, Portugal

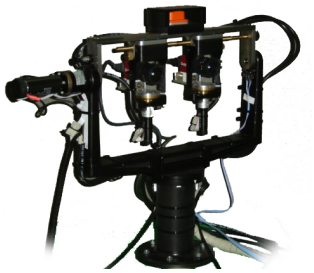


Figure 1. View of the current version of the Integrated Multimodal Perception Experimental Platform (IMPEP). The active perception head mounting hardware and motors were designed by the Perception on Purpose (POP - EC project number FP6-IST-2004-027268) team of the ISR/FCT-UC, and the sensor systems mounted at the Mobile Robotics Laboratory of the same institute, within the scope of the Bayesian Approach to Cognitive Systems project (BACS - EC project number FP6-IST-027140).

A. Contributions of this work

In this text, we intend to present an integrated account of our work, which has been partially documented in previous publications [7–13], together with unpublished results. We will start in section II by presenting a bioinspired perceptual solution with focus on Bayesian visuoauditory integration. This solution serves as a short-term spatial memory framework for active perception and also sensory control of action, with no immediate interest in object perception. We will try to explicitly or implicitly address each of the challenges of sensor fusion as described by Ernst and Bülthoff [14] using the *Bayesian Volumetric Map* (BVM) framework for vision, audition and vestibular sensing. It is our belief that perceptual systems are unable to yield useful descriptions of their environment without resorting to a temporal series of sensory fusion processed on some kind of short-term memory such as the BVM.

We mainly expect to contribute with a solution which:

- Deals inherently with perceptual uncertainty and ambiguity.
- Deals with the geometry of sensor fusion in a natural fashion.
- Allows for fast processing of perceptual inputs to build a spatial representation that promotes immediate action on the environment.

The Bayesian models for visuoauditory perception which form the backbone of the framework are presented next in section II. We propose to use proprioception (e.g. vestibular sensing) as ancillary information to promote visual and auditory sensing to satisfy the requirements for integration.

To support our research work and to provide a test-bed for some of the possible applications of the BVM, an artificial multimodal perception system (IMPEP — Integrated Multimodal Perception Experimental Platform) has been constructed at the ISR/FCT-UC. The platform consists 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 Fig. 1. It presents the same sensory capabilities as a human head, thus conforming to the biological conditions that originally inspired our work. We believe IMPEP has great potential for use in applications as diverse as active perception in social robots or even robotic navigation. We present a brief description of its implementation, its sensory processing modules and system calibration in section III.

As an illustration of the particular application of active perception, and also and more importantly to test the performance of our solution, in section IV we will present an algorithm that implements an active exploration behaviour based on the entropy of the BVM framework, together with results of using this algorithm in real-time.

Finally, in section V conclusions will be drawn and related ongoing work will be mentioned, and in section VI future work based on what is presented in this article will be discussed.

B. Related work

Fusing computer vision, binaural sensing and vestibular sensing using a unified framework, to the authors' knowledge, has never been addressed. Moreover, as far as is known by the authors, the application of the well-known probabilistic inference grid model [15] to an egocentric, log-spherical spatial configuration as a solution to problems remotely similar to the ones presented in this text is also unprecedented.

In our specific application domain, where a 3D metric and egocentric representation is required, common inference grid configurations which assume regularly partitioned Euclidean space to build the cell lattice are not appropriate:

1) Most sensors, vision and audition being notable examples, are based on a process of energy projection onto transducers, ideally yielding a pencil of projection lines that converge at the egocentric reference origin; consequently, they are naturally disposed to be directly modelled in polar or spherical coordinates. The only example of the use of a spherical configuration known to the authors was presented by Zapata et al. [16].
2) Implementation-wise, regular partitioning in Euclidean space, while still manageable in 2D, renders temporal performances impractical in 3D

when fully updating a panoramic grid (i.e. performing both prediction/estimation for **all** cells on the grid) with satisfactory size and resolution (typically grids with much more than a million cells). There are, in fact, two solutions for this problem: either non-regular partitioning of space (e.g. octree compression), or regular partitioning of log-distance space. Interestingly enough, the latter also accounts for just-discriminable depth thresholds found in human visual perception — an example of an Euclidean solution following a similar rationale was presented by Dankers, Barnes, and Zelinsky [17].

An important part of recent work in active vision, contrary to our solution, either use an explicit representation for objects to implement active perception or multisensory fusion (e.g. [18, 19]) or rely on object detection/recognition to establish targets for active object search (e.g. [20–22]). On the other hand, several solutions for applications similar to ours (e.g. [23–25]) avoid explicit object representation by resorting to a bottom-up saliency approach such as defined by Itti, Koch, and Niebur [26]. The underlying rationale is that postponing data association processing allows for the implementation of fast automatic mechanisms of active exploration that resemble what is believed to occur within the human brain, as explained in the introductory section of this paper.

Our solution implements active visuoauditory perception using an egocentric spatial representation, adding to it vestibular sensing/proprioception so as to allow for efficient sensor fusion given a rotational egomotion. Moreover, the log-partitioning of the spatial representation intrinsically deals with just-discriminable depth thresholds, while avoiding the use of complex error dispersion models. Complementing the possibility of the extension of our system to include sensory saliency-fuelled behaviours [13], our solution differs from purely saliency-based approaches in that it inherently implements an active exploration behaviour based on the entropy of the occupancy grid (inspired in research work such as [27]), so as to promote gaze shifts to regions of high uncertainty. In summary, our framework elicits an automatic behaviour of fixating interesting (i.e. salient) and unexplored regions of space, without the need to resort to active object search, as in [20–22].

II. Bayesian Models for Multimodal Perception

A. Background and definitions

Taking into account the goals stated in the introductory section, the framework for spatial representation that will be presented in the rest of this section satisfies the following criteria:

- It is egocentric and metric in nature.
- It is an inference grid, allowing for a probabilistic representation of dynamical spatial occupation of the environment. It therefore encompasses positioning, structure and motion of objects, avoiding any need for any assumptions on the nature of those objects, or in other words, for data association.

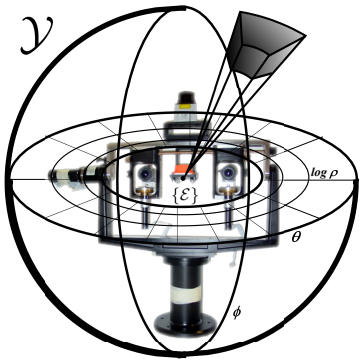


Figure 2. Egocentric, log-spherical configuration of the Bayesian Volumetric Map.

Given these requirements, we chose a *log-spherical* coordinate system spatial configuration (see Fig. 2) for the occupancy grid that we have developed and will refer to as BVM, thus promoting an egocentric trait in agreement with biological perception.

The BVM 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 $b = a^{\frac{\log_a(\rho_{\text{Max}} - \rho_{\text{Min}})}{N}}$, $\forall a \in \mathbb{R}$, where ρ_{Min} defines the *egocentric gap*, for a given number of partitions N, chosen according to application requirements. The BVM space is therefore effectively defined by

$$\mathcal{Y} \equiv]\log_b \rho_{\text{Min}}; \log_b \rho_{\text{Max}}] \times]\theta_{\text{Min}}; \theta_{\text{Max}}] \times]\phi_{\text{Min}}; \phi_{\text{Max}}] \quad (1)$$

In practice, the BVM is parametrised so as to cover the full angular range for azimuth and elevation. This configuration virtually delimits a *horopter* for sensor fusion.

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

$$\mathcal{Y} \supset \mathcal{C} \equiv]\log_b \rho_{\text{min}}; \log_b \rho_{\text{max}}] \times]\theta_{\text{min}}; \theta_{\text{max}}] \times]\phi_{\text{min}}; \phi_{\text{max}}] \quad (2)$$

Program

Description

Specification

Relevant variables:

$C \in \mathcal{Y}$: indexes a cell on the BVM;

A_C: identifier of the antecedents of cell C (stored as with C);

$Z_1, \cdots, Z_S \in \{\text{“No Detection”}\} \cup \mathcal{Z}$: *independent* measurements taken by S sensors;

O_C, O_C^{-1}: binary values describing the occupancy of cell C, for current and preceding instants, respectively;

V_C: velocity of cell C, discretised into $N+1$ possible cases $\in \mathcal{V} \equiv \{v_0, \cdots, v_N\}$.

Decomposition:

$$P(C\ A_C\ O_C\ O_C^{-1}\ V_C\ Z_1 \cdots Z_S) =$$

$$P(A_C)P(V_C|A_C)P(C|V_C\ A_C)P(O_C^{-1}|A_C)P(O_C|O_C^{-1}) \prod_{i=1}^{S} P(Z_i|V_C\ O_C\ C)$$

Parametric forms:

$P(A_C)$: uniform;

$P(V_C|A_C)$: histogram;

$P(C|V_C\ A_C)$: Dirac, 1 *iff* $c_{\log_b \rho} = a_{\log_b \rho} + v_{\log_b \rho}\delta t$, $c_\theta = a_\theta + v_\theta \delta t$ and $c_\phi = a_\phi + v_\phi \delta t$ (*constant velocity assumption*);

$P(O_C^{-1}|A_C)$: probability of preceding state of occupancy given set of antecedents;

$P(O_C|O_C^{-1})$: defined through transition matrix $T = \left[\begin{smallmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{smallmatrix}\right]$, where ϵ represents the probability of non-constant velocity;

$P(Z_i|V_C\ O_C\ C)$: *direct measurement model* for each sensor i, given by respective sub-BP.

Identification:

None.

Questions:

$$P(O_c\ V_c|z_1 \cdots z_S\ c) \rightarrow \begin{cases} P(O_c|z_1 \cdots z_S\ c) \\ P(V_c|z_1 \cdots z_S\ c) \end{cases}$$

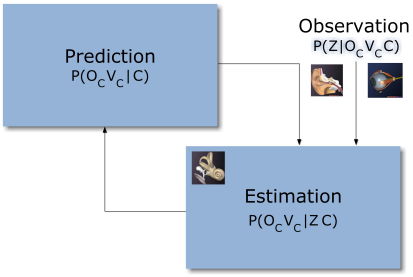


Figure 3. Bayesian Program for the estimation of Bayesian Volumetric Map current cell state (on the left), and corresponding Bayesian filter diagram (on the right – it considers only a single measurement Z for simpler reading, with no loss of generality). The respective filtering equation is given by (3) and (4), using two different formulations.

$$\overbrace{P(V_C\ O_C\ Z_1 \cdots Z_S\ C)}^{\text{Estimation (Joint Distribution)}} = \overbrace{\prod_{i=1}^{S} P(Z_i|V_C\ O_C\ C)}^{\text{Observation}} \overbrace{\sum_{A_C, O_C^{-1}} P(A_C)P(V_C|A_C)P(C|V_C\ A_C)P(O_C^{-1}|A_C)P(O_C|O_C^{-1})}^{\text{Prediction}} \tag{3}$$

$$\overbrace{P(V_C\ O_C|Z_1 \cdots Z_S\ C)}^{\text{Estimation}} = \frac{\overbrace{\prod_{i=1}^{S} P(Z_i|V_C\ O_C\ C)}^{\text{Observation}} \overbrace{\sum_{A_C, O_C^{-1}} P(A_C)P(V_C|A_C)P(C|V_C\ A_C)P(O_C^{-1}|A_C)P(O_C|O_C^{-1})}^{\text{Prediction}}}{\underbrace{\sum_{A_C, O_C^{-1}, O_C, V_C} P(A_C)P(V_C|A_C)P(C|V_C\ A_C)P(O_C^{-1}|A_C)P(O_C|O_C^{-1}) \prod_{i=1}^{S} P(Z_i|V_C\ O_C\ C)}_{\text{Normalisation}}} \tag{4}$$

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

The main hypothesis of inference grids is that the state of each cell is considered independent of the states of the remaining cells on the grid. This assumption effectively breaks down the complexity of state estimation. As a matter of fact, complete estimation of the state of the grid resumes to applying N times the cell state estimation process, N being the total number of cells that compose the grid.

To compute the probability distributions for the current states of each cell, the *Bayesian Program* (BP) formalism, consolidated by Bessière et al. [28], will be used throughout this text.

B. Multimodal Sensor Fusion Using Log-Spherical Bayesian Volumetric Maps

1) Using Bayesian filtering for visuoauditory integration: The independency hypothesis postulated earlier allows for the independent processing of each cell, and hence the Bayesian Program should be able to perform the evaluation of the state of a cell knowing an observation of a particular sensor.

The Bayesian Program presented in Fig. 3 is based on the solution presented by Tay et al. [29], the *Bayesian Occupancy Filter* (BOF), adapted so as to conform to the BVM egocentric, three-dimensional and log-spherical nature. In the spirit of Bayesian programming, we start by stating and defining the relevant variables:

- $C \equiv (\log_b \rho_{\max}, \theta_{\max}, \phi_{\max}) \in \mathcal{Y}$ is random variable denoting a log-spherical index which simultaneously localises and identifies the reference BVM cell, as has been defined in section II-A. It is used as a subscript of most of the random variables defined in this text, so as to explicitly state their relation to cells in the grid.
- $A_C \equiv (\log_b \rho_{\max}, \theta_{\max}, \phi_{\max}) \in \mathcal{A}_C \subset \mathcal{Y}$ is a random variable that denotes the *hypothetical antecedent* cell of reference cell C. The set of allowed antecedents $\mathcal{A}_C$ of reference cell C is composed by the $N+1$ cells on the BVM grid from which an object might have moved from, within the time interval going from the previous inference step $t-1$ to the present time t. The number of possible antecedents of any cell is arbitrary; in the case of the present work, we considered $N+1=7$ antecedents: two immediate neighbours in distance, two immediate neighbours in azimuth, and two immediate neighbours in elevation, and cell C itself (which would represent the hypothesis of an object occupying the reference cell remaining still).
- O_C is a binary variable denoting the occupancy $[O_C = 1]$ or emptiness $[O_C = 0]$ of cell C; O_C^{-1} denotes the occupancy state of the effective antecedent of C, A_C, in the previous inference step, which will propagate to the reference cell as the object occupying a specific A_C is moved to C.
- V_C denotes the dynamics of the occupancy of cell C as a vector signalling local motion to this cell from its antecedents, discretised into $N+1$ possible cases for velocities $\in \mathcal{V} \equiv \{v_0, \cdots, v_N\}$, with v_0 signalling that the most probable antecedent of A_C is C, i.e. no motion between two consecutive time instants.
- $Z_1, \cdots, Z_S \in \{\text{“No Detection”}\} \cup \mathcal{Z}$ are *independent* measurements taken by S sensors.

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. 3. This inference effectively leads to the Bayesian filtering formulation as used in the BOF grids.

Using the decomposition equation given in Fig. 3, we also have a more familiar formulation of the Bayesian filter of (3), given that $\prod_{i=1}^{S} P(Z_i|V_C\, O_C\, C)$ does not depend either on A_C or O_C^{-1}. Applying marginalisation and Bayes rule, we obtain the answer to the Bayesian Program question, the global filtering equation (4).

The process of solving the global filtering equation can actually be separated into three stages, in practice. The first stage consists on the prediction of the probabilities of each occupancy and velocity state for cell $[C = c]$, $\forall k \in \mathbb{N}_0, 0 \leq k \leq N$,

$$\alpha_c([O_C = 1], [V_C = v_k]) = \sum_{A_C, O_C^{-1}} P(A_C) P(v_k|A_C) P(C|v_k\, A_C) P(O_C^{-1}|A_C) P(o_c|O_C^{-1}) \tag{5a}$$

$$\alpha_c([O_C = 0], [V_C = v_k]) = \sum_{A_C, O_C^{-1}} P(A_C) P(v_k|A_C) P(C|v_k\, A_C) P(O_C^{-1}|A_C) P(\neg o_c|O_C^{-1}) \tag{5b}$$

with o_c and $\neg o_c$ used as shorthand notations for $[O_C = 1]$ and $[O_C = 0]$, respectively.

The prediction step thus consists on performing the computations represented by (5) for each cell, essentially by taking into account the velocity probability $P([V_C = v_k]|A_C)$ and the occupation probability of the set of antecedent cells represented by $P(O_C^{-1}|A_C)$, therefore propagating occupancy states as a function of the velocities of each cell.

The second stage of the BVM Bayesian filter estimation process is multiplying the results given by the previous step with the observation from the sensor model, yielding, $\forall k \in \mathbb{N}_0, 0 \leq k \leq N$,

$$\beta_c([O_C = 1], [V_C = v_k]) = \prod_{i=1}^{S} \big(P(Z_i|v_k\, [O_C = 1]\, C)\big)\ \alpha_c([O_C = 1], v_k) \tag{6a}$$

$$\beta_c([O_C = 0], [V_C = v_k]) = \prod_{i=1}^{S} \big(P(Z_i|v_k\, [O_C = 0]\, C)\big)\ \alpha_c([O_C = 0], v_k) \tag{6b}$$

Performing these computations for each cell $[C = c]$ gives a non-normalised estimate for velocity and occupancy for each cell. The marginalisation over occupancy values gives the likelihood of each velocity,

$$\forall k \in \mathbb{N}_0, 0 \leq k \leq N,$$

$$l_c(v_k) = \beta_c([O_C = 1], [V_C = v_k]) + \beta_c([O_C = 0], [V_C = v_k]) \quad (7)$$

The final normalised estimate for the joint state of occupancy and velocity for cell $[C = c]$ is given by

$$P(O_C\,[V_C = v_k]|Z_1 \cdots Z_S\,C) = \frac{\beta_c(O_C,[V_C=v_k])}{\sum_{V_C} l_c(V_C)} \quad (8)$$

The related remaining questions of the BP for the BVM cell states, the estimation of the probability of occupancy and the estimation of the probability of a given velocity, are given through marginalisation of the free variable by

$$P(O_C|Z_1 \cdots Z_S\,C) = \sum_{V_C} P(V_C\,O_C|Z_1 \cdots Z_S\,C) \quad (9a)$$

$$P(V_C|Z_1 \cdots Z_S\,C) = \sum_{O_C} P(V_C\,O_C|Z_1 \cdots Z_S\,C) \quad (9b)$$

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

2) 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 axis and the whole perceptual system is not allowed to perform any translation. In this case, the vestibular sensor models, described ahead, will yield measurements of angular velocity and position. These can then be easily used to manipulate the BVM, which is, by definition, in spherical coordinates.

Therefore, to compensate for this kind of *egomotion*, 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* $C = (\log_b \rho_{max}, \theta_{max} + \theta_{inertial}, \phi_{max} + \phi_{inertial})$.

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

3) Dealing with sensory synchronisation: The BVM model presented earlier assumes that the state of a cell C, given by (O_C, V_C), and the observation by any sensor i, given by Z_i, correspond to the same time instant t.

In accordance with the wide multisensory integration temporal window theory for human perception reviewed in [30], the BVM may be used safely to integrate auditory and vision measurements as soon they become available; local motion estimation using the BVM enforces a periodical state update with constant rate to ensure temporal consistency. Consequently, the modality of highest measurement rate is forced to set the update pace (i.e. by means of measurement buffers) in order to satisfy the constant update requirement. The velocity estimates for the local motion states of the BVM are thus a function of this update rate.

Promotion through vestibular sensing is also perfectly feasible, since inertial readings are available at a much faster rate than visuoauditory perception.

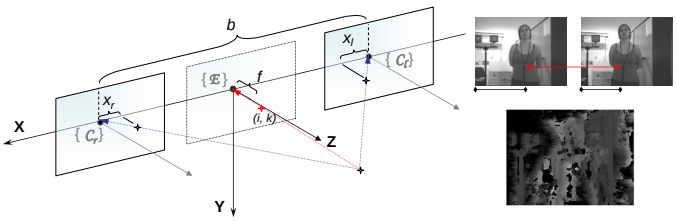


Figure 4. Cyclopean geometry for stereovision — b stands for the baseline and f is the focal length. The use of cyclopean geometry (pictured on the left for an assumed frontoparallel configuration) allows direct use of the egocentric reference frame for depth maps taken from the disparity maps yielded by the stereovision system (of which an example is shown on the right).

C. Bayesian sensor models

Next, the sensor models that are used as observations for the Bayesian filter of the BVM will be presented. C as a random variable and $P(C)$, although redundant in this context, will be used in the following models to maintain consistency with the Bayesian filter formulation and also with cited work.

1) Vision sensor model: 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, Aycard, and Laugier [31]. This solution incorporates a complete formal definition of the physical phenomenon of occlusion (i.e. in the case of visual occlusion, light reflecting from surfaces occluded by opaque objects do not reach the vision sensor's photoreceptors).

Our motivations suggest a tentative data structure analogous to neuronal population activity patterns to represent uncertainty in the form of probability distributions [32]. Thus, a spatially organised 2D grid may have each cell (corresponding to a virtual photoreceptor in the cyclopean view — see Fig. 4) associated to a "population code" extending to additional dimensions,

yielding a set of probability values encoding a N-dimensional probability distribution function or pdf.

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

$$P_k(Z) = L_k(Z, \mu_\rho(k), \sigma_\rho(k)), \begin{cases} \mu_\rho(k) & = \hat{\rho}(\hat{\delta}) \\ \sigma_\rho(k) & = \frac{1}{\lambda}\sigma_{min} \end{cases}, \qquad (10)$$

a discrete probability distribution with mean μ_ρ and standard deviation σ_ρ, both a function of the cell index k, which directly relates to the log-distance from the observer ρ. Values $\hat{\delta}$ and λ represent the disparity reading and its correspondent confidence rating, respectively; σ_{min} and the expression for $\hat{\rho}(\hat{\delta})$ are 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 (see camera calibration description later in this text). The likelihood function *constitutes, in fact, the elementary sensor model* as defined above for each vision sensor, and formally represents *soft evidence*, or "Jeffrey's evidence" in reference to Jeffrey's rule [33] concerning the relation between vision sensor measurements denoted generically by Z and the corresponding readings $\hat{\delta}$ and λ, described by the calibrated expected value $\hat{\rho}(\hat{\delta})$ and standard deviation $\sigma_\rho(\lambda)$ for each sensor.

Equation (10) only partially defines the resulting probability distribution by specifying the random variable over which it is defined and an expected value plus a standard deviation/variance — a full definition requires the choice of a type of distribution that best fits the noisy pdfs taken from the population code data structure. The traditional choice, mainly due to the central limit theorem, favours normal distributions $\mathcal{N}(Z, \mu_\rho(k), \sigma_\rho(k))$. Considering what happens in the mammalian brain, this choice appears to be naturally justified — biological population codes often yield bell-shaped distributions around a preferred reading [34, 35]. For more details on our adaptation of such a distribution, please refer to [7].

To correctly formalise the Bayesian inference process, a formal auxiliary definition with respective properties is needed — for details, please refer to [7]. The Bayesian Program that summarises this model is presented on Fig. 5.

2) Audition sensor model: Current trends in robotic implementations of sound-source localisation models rely on microphone arrays with more than a couple of sensors, either by resorting to steerable beamformers, high-resolution spectral estimation, time difference of arrival (TDOA) information, or fusion methods (i.e, the

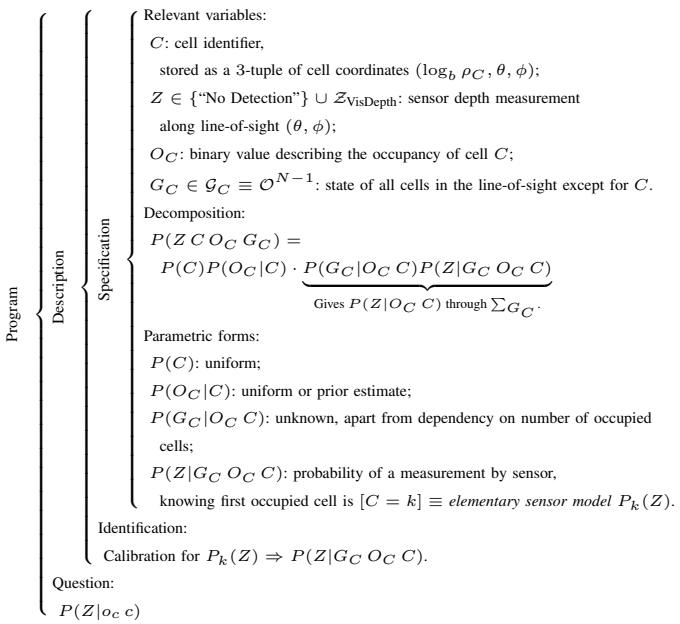


Figure 5. Bayesian Program for vision sensor model of occupancy.

integration of estimates from several distributed arrays). Generically, it is found that increasing the number of microphones also increases estimation accuracy. In fact, there is theoretical (see the Cramer-Rao bound analyses presented by Yang and Scheuing [36] for TDOA and Chen et al. [37] for beamforming) and practical (see Loesch et al. [38]) evidence supporting this notion. Conversely, it is also generally accepted that the computational burden of sound-source localisation increases with the number of sensors involved; in fact, this provides the support for the use of fusion methods, and also for one of the reasons speculated for why humans, like many animals, have only two ears [39].

The direct audition sensor model used in this work, first presented in [8, 10], relies on binaural cues alone to fully localise a sound-source in space. The reason for its use resides in our biological inspiration mentioned in section I-A and our desire to use the BVM framework in future experiments where comparisons to human sensory systems are to be made in "fair terms" (see future work referred to in [13]); however, the inclusion within this framework of any alternative model supporting the use of more microphones to increase estimation accuracy would be perfectly acceptable. The model is formulated as the first question of the Bayesian Program in Fig. 6, where all relevant variables and distributions and the decomposition of the corresponding joint distribution, according to Bayes' rule and dependency assumptions, are defined. The use of the auxiliary binary random variable S_C, which signals the presence or absence of a sound-source

Program

Description

Specification

Relevant variables:

$C \equiv (\log_b \rho_{\max}, \theta_{\max}, \phi_{\max}) \in \mathcal{C}$: cell identifier;

$Z \in \mathcal{Z}_{\text{BinauralMeasurements}}$: sensor measurement vectors $[\tau, \Delta L(f_c^1) \cdots \Delta L(f_c^m)]$;

$\tau \equiv$ ITD and $\Delta L(f_c^k) \equiv$ ILD; f_c^k denotes each $k \in \mathbb{N}, 1 \leq k \leq m$ frequency band in m frequency channels).

S_C: binary value describing the presence of a sound-source in cell C, $[S_C = 1]$ if a sound-source is present at C, $[S_C = 0]$ otherwise;

O_C: binary value describing the occupancy of cell C, $[O_C = 1]$ if cell C is occupied by an object, $[O_C = 0]$ otherwise;

Decomposition:

$$P(Z\, C\, S_C\, O_C) = P(C)P(O_C|C)\cdot \underbrace{P(S_C|O_C\, C)P(\tau|S_C\, O_C\, \theta_{\max}) \prod_{k=1}^{m} P(\Delta L(f_c^k)|\tau\, S_C\, O_C\, C)}_{\text{Gives } P(Z|O_C\, C) \text{ through } \sum_{S_C}}$$

Parametric forms:

$P(C)$: uniform;

$P(O_C|C)$: uniform or prior estimate;

$P(S_C|O_C\, C) \equiv P(S_C|O_C)$: probability table (see table below);

$P(Z|O_C\, C)$: probability of a measurement $[\tau, \Delta L(f_c^1) \cdots \Delta L(f_c^m)]$;

$P(\tau|S_C\, O_C\, \theta_{\max}) \equiv P(\tau|S_C\, \theta_{\max})$: normal distribution, yielding the probability of a measurement τ by sensor for cell C, given its azimuth $\theta_{\max}$ and presence or absence of a sound-source S_C in that cell;

$P(\Delta L(f_c^k)|\tau\, S_C\, O_C\, C) \equiv P(\Delta L(f_c^k)|\tau\, S_C\, C)$: normal distribution, yielding the probability of a measurement $\Delta L(f_c^k)$ by sensor for cell C, given the presence or absence of a sound-source S_C in that cell.

Identification:

Calibration for $P(\tau|S_C\, O_C\, \theta_{\max})$.

Calibration for $P(\Delta L(f_c^k)|\tau\, S_C\, O_C\, C) \approx P(\Delta L(f_c^k)|S_C\, O_C\, C)$.

Questions:

$P(Z|o_c\, c)$

$\max, \arg\max_C P([S_C = 1]|z\, C)$

$P(S_C\|O_C)$	$[O_C = 0]$	$[O_C = 1]$
$[S_C = 0]$	1	.5
$[S_C = 1]$	0	.5
$\sum P(s_c\|O_C)$	1	1

Figure 6. Bayesian Program for binaural sensor model. At the bottom is presented the probability table which was used for $P(S_C|O_C\, C) \equiv P(S_C|O_C)$, empirically chosen so as to reflect the indisputable fact that there is no sound-source in a cell that is not occupied (left column), and the safe assumption that when a cell is known to be occupied there is little way of telling if it is in this condition due to a sound-source or not (right column).

in cell C, and the corresponding family of probability distributions $P(S_C|O_C\, C) \equiv P(S_C|O_C)$ promotes the assignment of probabilities of occupancy close to 1 for cells for which the binaural cue readings seem to indicate a presence of a sound-source and close to .5 otherwise (i.e. the absence of a detected sound-source in a cell doesn't mean that the cell is empty). The second question corresponds to the estimation of the position of cells most probably occupied by sound-sources, through the inversion of the direct model through Bayesian inference on the joint distribution decomposition equation.

The former is used as a sub-BP for the BVM multi-

Program

Description

Specification

Relevant variables:

$\boldsymbol{\xi}^t = (\boldsymbol{\Theta}^t, \boldsymbol{\Omega}^t, \boldsymbol{A}^t)$: state variables,

$\boldsymbol{S}^t = (\boldsymbol{\Phi}^t, \boldsymbol{\Upsilon}^t)$: sensor variables.

Decomposition:

$$P(\boldsymbol{\xi}^t\, \boldsymbol{\xi}^{t-\delta t}\, \boldsymbol{S}^t \ldots\, \boldsymbol{S}^0) = P(\boldsymbol{S}^t|\boldsymbol{\xi}^t) .P(\boldsymbol{\Omega}^t).P(\boldsymbol{A}^t).P(\boldsymbol{\Theta}^t|\boldsymbol{\Theta}^{t-\delta t}\, \boldsymbol{\Omega}^t) .P(\boldsymbol{\xi}^{t-\delta t}\, \boldsymbol{S}^{t-\delta t} \ldots\, \boldsymbol{S}^0)$$

Parametric forms:

$P(\boldsymbol{S}^t|\boldsymbol{\xi}^t) = P(\boldsymbol{\Phi}^t|\boldsymbol{\Omega}^t).P(\boldsymbol{\Upsilon}^t|\boldsymbol{F}^t).P(\boldsymbol{F}^t|\boldsymbol{\Theta}^t\, \boldsymbol{A}^t)$: sensor model, Gaussians and dirac;

$P(\boldsymbol{\Omega}^t)$, $P(\boldsymbol{A}^t)$: a priori for state, Gaussians;

$P(\boldsymbol{\Theta}^t|\boldsymbol{\Theta}^{t-\delta t}\, \boldsymbol{\Omega}^t)$: state dynamic model, diracs;

$P(\boldsymbol{\xi}^{t-\delta t}\, \boldsymbol{S}^{t-\delta t} \ldots\, \boldsymbol{S}^0)$: previous iteration, distribution computed at last time step.

Identification:

Parameters of the Gaussians: σ_Φ, σ_Υ, σ_A and σ_Ω .

Question:

$P(\boldsymbol{\xi}^t|\boldsymbol{S}^t \ldots\, \boldsymbol{S}^0)$

Figure 7. Bayesian Program for processing of inertial data.

modal sensor fusion framework described earlier, while the answer to the latter yields a gaze direction of interest in terms of auditory features which can be used by a multimodal attention system, through a maximum *a posteriori* (MAP) method.

3) Vestibular sensor model: To process the inertial data, we follow the Bayesian model of the human vestibular system proposed by Laurens and Droulez [40, 41], adapted here to the use of inertial sensors. The aim is to provide an estimate for the current angular position and angular velocity of the system, that mimics the human vestibular perception.

At time t the Bayesian program of Fig. 7 allows the inference of the probability distribution of the current state $\boldsymbol{\xi}^t = (\boldsymbol{\Theta}^t, \boldsymbol{\Omega}^t, \boldsymbol{A}^t)$ — where the orientation of the system in space is encoded using a rotation matrix $\boldsymbol{\Theta}$, the instantaneous angular velocity is defined as the vector $\boldsymbol{\Omega}$, and linear acceleration by $\boldsymbol{A}$ — given all the previous sensory inputs until the present instant, represented by $\boldsymbol{S}^{0\to t} = (\boldsymbol{\Phi}^{0\to t}, \boldsymbol{\Upsilon}^{0\to t})$ — where $\boldsymbol{\Phi}$ denotes $\boldsymbol{\Omega}$ with added Gaussian noise measured by the gyros, and $\boldsymbol{\Upsilon}$ denotes the gravito-inertial acceleration $\boldsymbol{F}$ with added Gaussian noise measured by the accelerometers — and the initial distribution $\boldsymbol{\xi}^0$. Details regarding this model, the respective implementation issues and preliminary results have been presented in [11, 12].

III. The Integrated Multimodal Perception Experimental Platform

A. Platform description

To support our research work, we have developed an artificial multimodal perception system, of which an im-

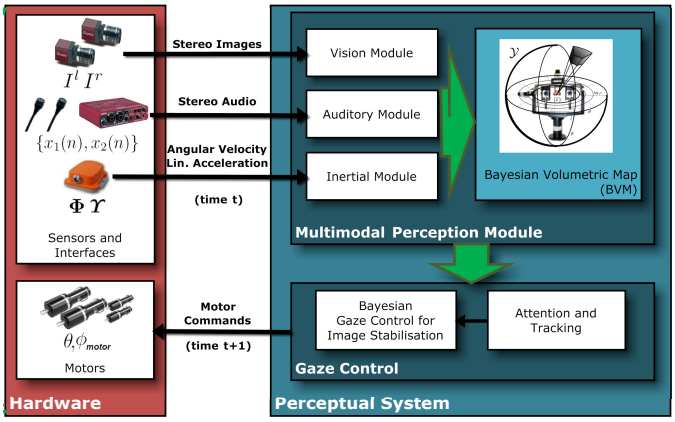


Figure 8. Implementation diagram for the BVM-IMPEP multimodal perception framework.

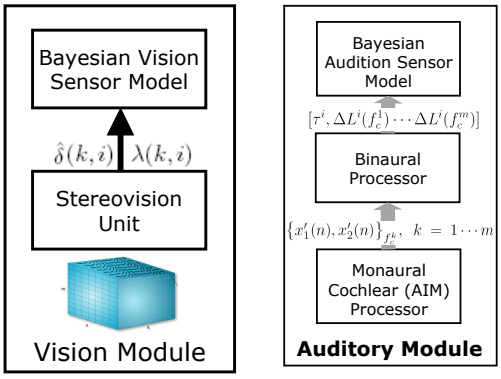


Figure 9. The IMPEP Bayesian sensor systems.

plementation diagram is presented on Fig. 8, 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.

The stereovision system is implemented using a pair of Guppy IEEE 1394 digital cameras from Allied Vision Technologies (http://www.alliedvisiontec.com), the binaural setup using two AKG Acoustics C417 linear microphones (http://www.akg.com/) and an FA-66 Firewire Audio Capture interface from Edirol (http://www.edirol.com/), and the miniature inertial sensor, Xsens MTi (http://www.xsens.com/), provides digital output of 3D acceleration, 3D rate of turn (rate gyro) and 3D earth-magnetic field data for the IMU.

Full implementation details can be found in [12].

B. Sensory processing

In the following text, the foundations of the sensory processing systems depicted on Fig. 9, which feed the Bayesian sensor models that have been defined in previous text, will be presented.

1) Vision system: The stereovision algorithm used yields an estimated disparity map $\hat{\delta}(k,i)$ and a corresponding confidence map $\lambda(k,i)$, and is thus easily converted from its deterministic nature into a probabilistic implementation simulating the population code-type data structure, as defined earlier.

2) Auditory system: The Bayesian binaural system presented herewith is composed of three distinct and consecutive modules (Fig. 9): 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 localisation unit*, which applies a Bayesian sensor model so as to perform localisation of sound-sources in 3D space. We have adapted the realtime software by the Speech and Hearing Group at the University of Shefield [42] to implement the solution by Faller and Merimaa [43] as the binaural processor — for more details, please refer to [8, 10].

3) Inertial sensing system: The calibrated inertial sensors in the IMU provide direct egocentric measurements of body angular velocity and linear acceleration. The gyros and the accelerometers provide noise-corrupted measurements of angular velocity $\boldsymbol{\Omega}^t$ and the gravito-inertial acceleration $\boldsymbol{F}$ as input for the sensor model of Fig. 7.

C. System calibration

Camera calibration was performed using the Camera Calibration Toolbox by Bouguet [44], therefore allowing the application of the reprojection equation:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & f \\ 0 & 0 & \frac{1}{b} & 0 \end{bmatrix} \begin{bmatrix} u_l - \frac{\hat{\delta}}{2} \\ v_l \\ \hat{\delta} \\ 1 \end{bmatrix} = \begin{bmatrix} WX \\ WY \\ WZ \\ W \end{bmatrix} \tag{11}$$

where u_l and v_l represent the horizontal and vertical coordinates of a point on the left camera, respectively, and $\hat{\delta}$ is the disparity estimate for that point, all of which in pixels, f and b are the estimated focal length and baseline, respectively, both of which in metric distance, and X, Y and Z are 3D point coordinates respective to the egocentric/cyclopean referential system $\{\mathcal{E}\}$.

Using reprojection error measurements given by the calibration procedure, parameter σ_{min} as defined earlier

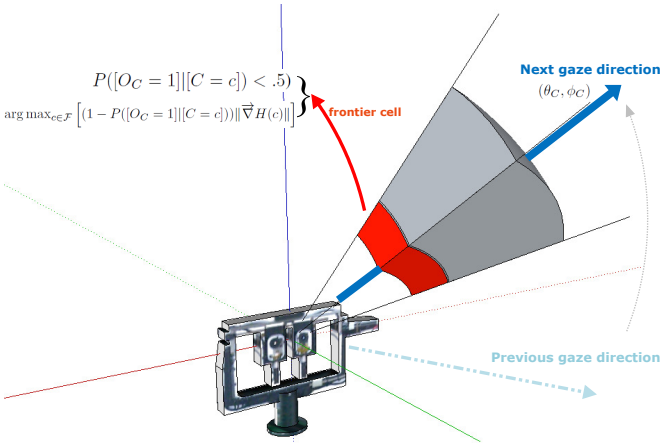


Figure 10. Illustration of the entropy-based active exploration process using the Bayesian Volumetric Map. Please refer to [9, 12] for more details.

is taken as being equal to the maximum error exhibited by the stereovision system.

Calibration of the binaural system involves the characterisation of the families of normal distributions $P(\tau|S_C\, O_C\, \theta_{\max})$ and $P(\Delta L(f_c^k)|\tau\, S_C\, O_C\, C) \approx P(\Delta L(f_c^k)|S_C\, O_C\, C)$ of the binaural sensor model defined earlier through descriptive statistical learning of their central tendency and statistical variability. This is done following a proceeding similar to commonly used head-related transfer function (HRTF) calibration processes, and was described in detail in [10].

Visuoinertial calibration was performed using the InerVis toolbox [45], that adds on to the Camera Calibration Toolbox by Bouguet [44]. The toolbox estimates the rotation quaternion between the Inertial Measurement Unit and a chosen camera, requiring a set of static observations of a vertical chequered visual calibration target and of sensed gravity [46].

IV. Active Exploration Using Bayesian Models for Multimodal Perception

A. Active exploration using the Bayesian Volumetric Map

Information in the BVM is stored as the *probability of each cell being in a certain state*, defined 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:

$$H(c) \equiv H(V_c, O_c) = -\sum_{\substack{o_c \in \mathcal{O} \\ v_c \in \mathcal{V}}} P(v_c\, o_c|z\, c) \log P(v_c\, o_c|z\, c) \quad (12)$$

The joint entropy value $H(c)$ is a sample of a continuous joint entropy field $H : \mathcal{Y} \to \mathbb{R}$, taken at log-spherical positions $[C = c] \in \mathcal{Y}$. 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

$$\vec{\nabla} H(c) \approx [H(c) - H(c_{\rho-}), H(c) - H(c_{\theta-}), H(c) - H(c_{\phi-})]^T \quad (13)$$

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

A great advantage of the BVM over Cartesian implementations of occupancy maps 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). In case there is more than one global joint entropy gradient maximum, the cell corresponding to the direction closest to the current heading is chosen, so as to deal with equiprobability, while simultaneously ensuring minimum gaze shift rotation effort (see Fig. 10). The full description of the active exploration heuristics was presented in [9, 12].

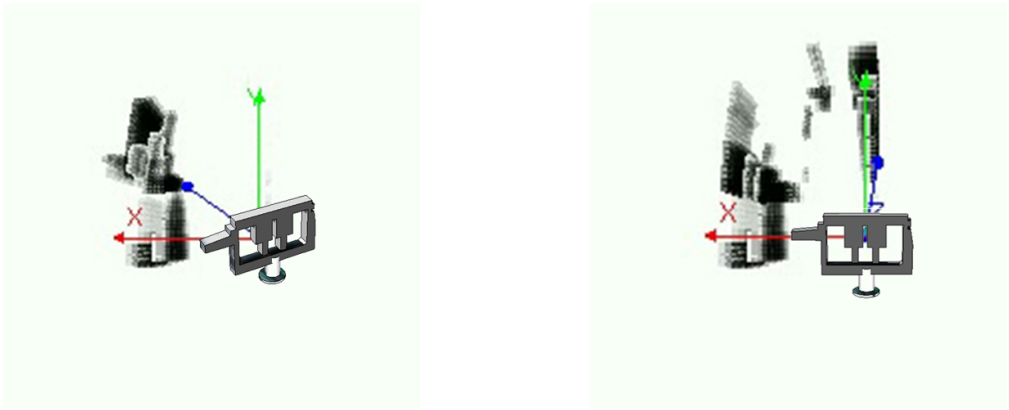
B. Results

The realtime implementation of all the processes of the framework was subjected to performance testing for each individual module — please refer to [12] for further details. To avoid Bayesian update deadlocks due to 0 or 1-probabilities, a simple error model analogous to what is proposed in [31] was implemented, both for occupancy and for velocities. Addtionally, log-probabilities are used to increase both the execution performance and the numerical stability of the system.

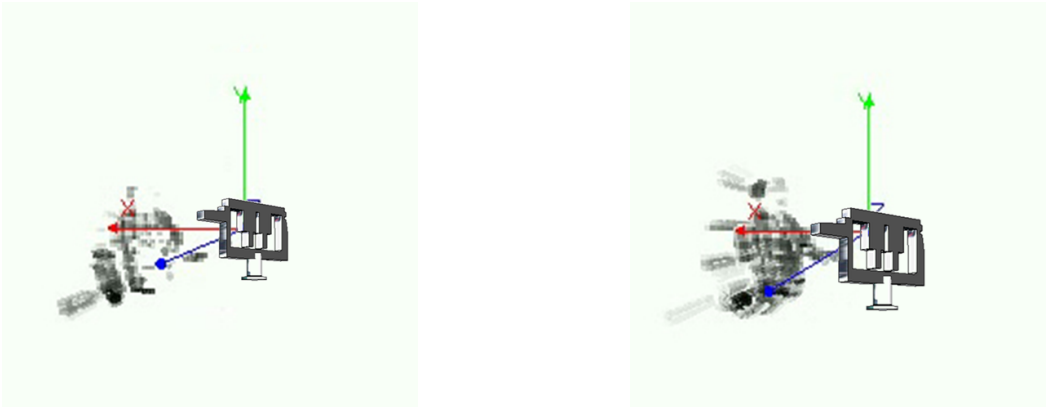
The full active exploration system runs at about 6 to 10 Hz. This is ensured by forcing the main BVM thread to pause for each time-step when no visual measurement is available (i.e. during 40 ms for $N = 10, \Delta\phi = 2^{\text{o}}$). This guarantees that BVM time-steps are regularly spaced, which is a very important requirement for correct implementation of prediction/dynamics, and also ensures that processing and memory resources are freed and unlocked regularly.

In Fig. 11 a qualitative comparison is made between the outcome of using each sensory modality individually, and also with the result of multimodal fusion, using a single speaker scenario, showcasing the advantages of visuoauditory integration in the effective use of both the spatial precision of visual sensing, and the temporal precision and panoramic capabilities of auditory sensing. The BVM renderings were produced from screenshots

(a) Left camera snapshot of a male speaker, at -41° azimuth relatively to the Z axis, which defines the frontal heading respective to the IMPEP "neck".



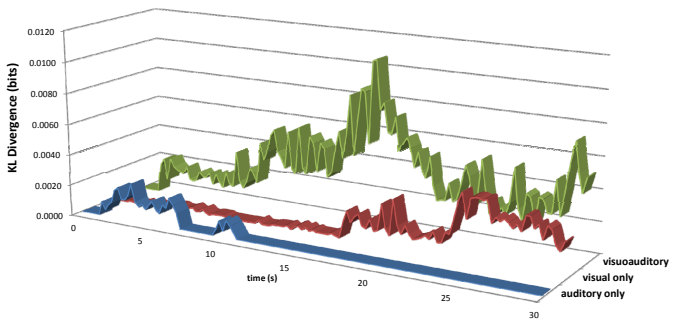
(b) BVM results for binaural processing only. Interpretation, from left to right: 1) sound coming from speaker triggers an estimate for occupancy from the binaural sensor model, and a consecutive exploratory gaze shift at approximately 1.6 seconds; 2) At approximately 10 seconds, noise coming from the background introduce a false positive, that is never again removed from the map (i.e. no sound does not mean no object, only no audible sound-source).



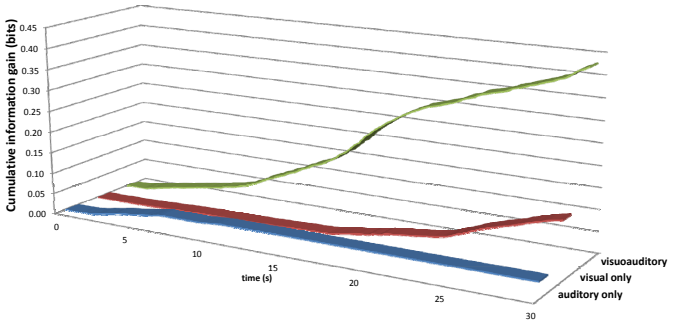
(c) BVM results for stereovision processing only. Notice the clean cut-out of speaker silhouette, as comparing to results in (b). On the other hand, active exploration using vision sensing alone took approximately 15 seconds longer to start scanning the speaker's position in space, while using binaural processing the speaker was fixated a couple of seconds into the experiment.

(d) BVM results for visuoauditory fusion. In this case, the advantages of both binaural (immediacy from panoramic scope) and stereovision (greater spatial resolution and the ability to clean empty regions in space) influence the final outcome of this particular instantiation of the BVM, taken at 1.5 seconds.

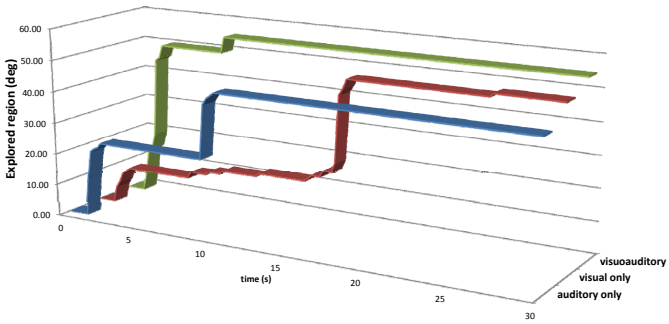
Figure 11. Online results for the real-time prototype for multimodal perception of 3D structure and motion using the BVM — three reenactments (binaural sensing only, stereovision sensing only and visuoauditory sensing) of a single speaker scenario. A scene consisting of a male speaker talking in a cluttered lab is observed by the IMPEP active perception system and processed online by the BVM Bayesian filter, using the active exploration heuristics described in the main text, in order to scan the surrounding environment. The heading arrow together with an oriented 3D sketch of the IMPEP perception system depicted in each map denote the current gaze orientation. All results depict frontal views, with Z pointing outward. The parameters for the BVM are as follows: $N = 10$, $\rho_{Min} = 1000\,\text{mm}$ and $\rho_{Max} = 2500\,\text{mm}$, $\theta \in [-180^{\circ}, 180^{\circ}]$, with $\Delta\theta = 1^{\circ}$, and $\phi \in [-90^{\circ}, 90^{\circ}]$, with $\Delta\phi = 2^{\circ}$, corresponding to $10 \times 360 \times 90 = 324,000$ 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 [6]).



(a) Instantaneous average information gain for updated cells.



(b) Cumulative sum of average information gain for updated cells.



(c) Maximum angle covered by explored volumetric convex hull.

Figure 12. Temporal evolution of average information gain (i.e. average Kullback-Liebler divergence for the full set of cells which were updated, either due to observations or propagation from prediction) and corresponding exploration span for the auditory-only, visual-only and visuoauditory versions of the single speaker scenario (see Fig. 11), for a 30 second period since the start of each experiment.

of an online OpenGL-based viewer running throughout the experiments. Fig. 12 presents a study based on information gain and exploration span, yielding a quantitative comparison of these advantages and capabilities, and demonstrating the superior results of visuoauditory fusion as compared to using each sensory modality separately. In [12], results were presented concerning the effectiveness of active exploration when having to deal with the ambiguity and uncertainty caused by multiple sensory targets and complex noise in a two-speaker scenario scenario. In order to provide a better view of the BVM configuration, the supplemental MPEG file presents a higher resolution rendering of a reenactment of this scenario, processed offline from a saved instantiation of the BVM occupancy grid.

The advantages of the log-spherical configuration were made apparent in these experiments, when comparing to other solutions in the literature: (a) an *efficiency advantage*: as mentioned in the introductory section, fewer cells are used to represent the environment — for example, to achieve the resolution equivalent to approximately the distance obtained half-way through the log-scale using a regular Euclidean partition for the examples in Figs. 11 and the supplemental video, i.e. 40 mm-side cells and removing the egocentric gap, around $1,937,500$ cells would be needed, while, using a similar rationale, around $1,215,000$ cells would be needed for a solution such as the one presented by [17], roughly constituting at least a 3-fold reduction in total cell count; (b) a *robustness advantage*: the fact that sensor readings are directly referred to in spherical coordinates and consequently no ray-tracing is needed leads to inherent antialiasing, therefore avoiding the Moiré effects which are present in other grid-based solutions in the literature, as is reported by Yguel et al. [31].

Moreover, the benefits of using an egocentric spherical configuration have also been made clear: sensory fusion is seamlessly performed, avoiding the consequences of transformations between referentials and respective complex registration and integration processes proposed in related work, such as [20–22].

V. Conclusions

In this text we introduced Bayesian models for visuoauditory perception and inertial sensing emulating vestibular perception which form the basis of the probabilistic framework for multimodal sensor fusion — the Bayesian Volumetric Map. These models build upon a common spatial configuration that is naturally fitting for

the integration of readings from multiple sensors. We also presented the robotic platform that supports the use of these computational models for implementing an entropy-based exploratory behaviour for multimodal active perception. In the future, the computational models described in this text will allow the construction of a simultaneously flexible and powerful robotic implementation of multisensory active perception to be used in real-world applications.

Regarding its future use in applications such as human-machine interaction or mobile robot navigation, the following conclusions may be drawn:

- The results presented in the previous section show that active exploration algorithm successfully drives the IMPEP-BVM framework to explore areas of the environment mapped with high uncertainty in real-time, with an intelligent heuristic that minimises the effects of local minima by attending to the closest regions of high entropy first.
- Moreover, since the human saccade-generation system promotes fixation periods (i.e. time intervals between gaze shifts) of a few hundred milliseconds on average [47, 48], the overall rates of 6 to $10\,\mathrm{Hz}$ achieved with our CUDA implementation, in our opinion, back up the claim that our system does, in fact, achieve satisfactory real-time performance.
- Effective use of visual spatial accuracy and auditory panoramic capabilities and temporal accuracy by our system constitutes a powerful solution for attention allocation in realistic settings, even in the presence of ambiguity and uncertainty caused by multiple sensory targets and complex noise.
- Although not explicitly providing for object representation, many of the scene properties that are already represented by the Bayesian filter allow for clustering and tracking of neighbouring cells sharing similar states, which in turn provides a fast processing prior/cue generator for an additional object detection and recognition module. An active object search could then be implemented, as in related work such as [20–22].

The BVM and its egocentric log-spherical configuration carry with it, however, in its current state, a few important limitations. In decreasing order of importance, these would be the following:

1) The non-regular tesselation of space might introduce perceptual distortions due to motion-based prediction when a moving object becomes occluded: a big object moving towards the observer will appear to shrink or, conversely, a small object moving away from the observer will appear to inflate. These perceptual illusions will of course disappear as soon as the object returns to the observer's line-of-sight.
2) If an object happens to get outside of the robotic observer's field-of-view, either due to a gaze shift or to object motion, the effect of the BVM representing a persistent memory might result in "ghost occupancies" and consequent cluttering of the spatial map. This did not visibly happen during the experiments described in section IV-B; however, it is a definite concern, and will be addressed in future work (see section VI).
3) If this system is to be used by an autonomous mobile robot to perform allocentric mapping, then some care must be taken when dealing with the non-trivial integration of reconstructions taken from different points of view. This is, however, beyond the scope of our current research.

We are currently developing a complex artificial active perception system that follows human-like bottom-up driven behaviours using vision, audition and vestibular sensing, building upon the work presented in this text. More specifically, we have devised a hierarchical modular probabilistic framework that allows the combination of active perception behaviours, adding to the active exploration behaviour described in this text automatic orientation based on sensory saliency. This research work has demonstrated in practice the usefulness rendered by the extensibility, adaptivity and scalability of the proposed framework – for more details, please refer to [13].

Further details on the development and application of these models can be found at http://paloma.isr.uc.pt/~jfilipe/BayesianMultimodalPerception.

VI. Future Work

Long-term improvements to the BVM-IMPEP framework would include sensor models specifically for local motion, in contrast to the occupancy-only-based sensor models presented in this paper. These models could be built upon concepts such as optical flow processing for vision (which could be enhanced by visuoinertial integration), the Doppler effect for audition, etc. – and perceptual grouping solutions, through clustering processes similar to what was presented by Tay et al. [29], but in our case using prior distributions based on multimodal perceptual integration processes, some of which are currently being studied in psychophysical studies performed by our research group, to be concluded soon.

Another important addition would be the introduction of a decay factor to the BVM – in other words a "forget-

fulness" factor – thus avoiding the cluttering limitation of the framework, pointed out in section V.

ACKNOWLEDGEMENTS

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

The authors would like to thank the reviewers and the Associate Editor for their kind and useful suggestions, which we gratefully acknowledge to have substantially improved the quality of our manuscript.

REFERENCES

[1] K. J. Murphy, D. P. Carey, and M. A. Goodale, "The Perception of Spatial Relations in a Patient with Visual Form Agnosia," *Cognitive Neuropsyshology*, vol. 15, no. 6/7/8, pp. 705–722, 1998.

[2] D. C. Knill and A. Pouget, "The Bayesian brain: the role of uncertainty in neural coding and computation," *TRENDS in Neurosciences*, vol. 27, no. 12, pp. 712–719, December 2004.

[3] M. J. Barber, J. W. Clark, and C. H. Anderson, "Neural representation of probabilistic information," *Neural Computation*, vol. 15, no. 8, pp. 1843–1864, August 2003.

[4] J. Gordon, M. F. Ghilardi, and C. Ghez, "Accuracy of planar reaching movements. I. Independence of direction and extent variability," *Experimental Brain Research*, vol. 99, no. 1, pp. 97–111, 1994.

[5] J. McIntyre, F. Stratta, and F. Lacquaniti, "Short-Term Memory for Reaching to Visual Targets: Psychophysical Evidence for Body-Centered Reference Frames," *Journal of Neuroscience*, vol. 18, no. 20, pp. 8423–8435, October 15 1998.

[6] J. E. Cutting and P. M. Vishton, "Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth," in *Handbook of perception and cognition*, W. Epstein and S. Rogers, Eds. Academic Press, 1995, vol. 5; Perception of space and motion.

[7] J. F. Ferreira, P. Bessière, K. Mekhnacha, J. Lobo, J. Dias, and C. Laugier, "Bayesian Models for Multimodal Perception of 3D Structure and Motion," in *International Conference on Cognitive Systems (CogSys 2008)*, University of Karlsruhe, Karlsruhe, Germany, April 2008, pp. 103–108.

[8] C. Pinho, J. F. Ferreira, P. Bessière, and J. Dias, "A Bayesian Binaural System for 3D Sound-Source Localisation," in *International Conference on Cognitive Systems (CogSys 2008)*, University of Karlsruhe, Karlsruhe, Germany, April 2008, pp. 109–114.

[9] J. F. Ferreira, C. Pinho, and J. Dias, "Active Exploration Using Bayesian Models for Multimodal Perception," in *Image Analysis and Recognition, Lecture Notes in Computer Science series (Springer LNCS), International Conference ICIAR 2008*, A. Campilho and M. Kamel, Eds., June 25–27 2008, pp. 369–378.

[10] ——, "Implementation and Calibration of a Bayesian Binaural System for 3D Localisation," in *2008 IEEE International Conference on Robotics and Biomimetics (ROBIO 2008)*, Bangkok, Thailand, February, 21–26 2009, pp. 1722–1727.

[11] J. Lobo, J. F. Ferreira, and J. Dias, "Robotic Implementation of Biological Bayesian Models Towards Visuo-inertial Image Stabilization and Gaze Control," in *2008 IEEE International Conference on Robotics and Biomimetics (ROBIO 2008)*, Bangkok, Thailand, February, 21–26 2009, pp. 443–448.

[12] J. F. Ferreira, J. Lobo, and J. Dias, "Bayesian real-time perception algorithms on GPU — Real-time implementation of Bayesian models for multimodal perception using CUDA," *Journal of Real-Time Image Processing*, vol. 6, no. 3, pp. 171–186, September 2011.

[13] J. F. Ferreira, M. Castelo-Branco, and J. Dias, "A hierarchical Bayesian framework for multimodal active perception," *Adaptive Behavior*, vol. 20, no. 3, pp. 172–190, June 2012.

[14] M. O. Ernst and H. H. Bülthoff, "Merging the senses into a robust percept," *TRENDS in cognitive Sciences*, vol. 8, no. 4, pp. 162–169, April 2004.

[15] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *IEEE Computer*, vol. 22, no. 6, pp. 46–57, 1989.

[16] R. Zapata, B. Jouvencel, and P. Lépinay, "Sensor-based motion control for fast mobile robots," in *IEEE Int. Workshop on Intelligent Motion Control*, Istambul, Turkey, 1990, pp. 451–455.

[17] A. Dankers, N. Barnes, and A. Zelinsky, "Active Vision for Road Scene Awareness," in *IEEE Intelligent Vehicles Symposium (IVS05)*, Los Vegas, USA, June 2005, pp. 187–192.

[18] J. Tsotsos and K. Shubina, "Attention and Visual Search : Active Robotic Vision Systems that Search," in *The 5th International Conference on Computer Vision Systems*, Bielefeld, March 21–24 2007.

[19] D. Roy, "Integration of Speech and Vision using Mutual Information," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 6, 2000, pp. 2369–2372.

[20] A. Andreopoulos, S. Hasler, H. Wersing, H. Janssen, J. K. Tsotsos, and E. Körner, "Active 3D Object Localization Using A Humanoid Robot," *IEEE Transactions on Robotics*, vol. 27, no. 1, pp. 47–64, 2011.

[21] J. Ma, T. H. Chung, and J. Burdick, "A probabilistic framework for object search with 6-DOF pose estimation," *International Journal of Robotics Research*, vol. 30, no. 10, pp. 1209–1128, 2011.

[22] F. Saidi, O. Stasse, K. Yokoi, and F. Kanehiro, "Online object search with a humanoid robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007)*, 2007, pp. 1677–1682.

[23] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati, "Active Vision for Sociable Robots," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 31, no. 5, pp. 443–453, September 2001.

[24] A. Dankers, N. Barnes, and A. Zelinsky, "A Reactive Vision System: Active-Dynamic Saliency," in *5th International Conference on Computer Vision Systems*, Bielefeld, Germany, 21–24 March 2007.

[25] A. Koene, J. Morén, V. Trifa, and G. Cheng, "Gaze shift reflex in a humanoid active vision system," in *5th International Conference on Computer Vision Systems*, Bielefeld, Germany, 2007.

[26] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[27] R. Rocha, J. Dias, and A. Carvalho, "Cooperative Multi-Robot Systems: a study of Vision-based 3-D Mapping using Information Theory," *Robotics and Autonomous Systems*, vol. 53, no. 3–4, pp. 282–311, December 2005.

[28] P. Bessière, C. Laugier, and R. Siegwart, Eds., *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, ser. Springer Tracts in Advanced Robotics. Springer, 2008, vol. 46, ISBN: 978-3-540-79006-8.

[29] C. Tay, K. Mekhnacha, C. Chen, M. Yguel, and C. Laugier, "An efficient formulation of the bayesian occupation filter for target tracking in dynamic environments," *International Journal of Autonomous Vehicles*, vol. 6, no. 1–2, pp. 155–171, 2008.

[30] C. Spence and S. Squire, "Multisensory integration: maintaining the perception of synchrony," *Current Biology*, vol. 13, pp. R519–R521, July 2003.

[31] M. Yguel, O. Aycard, and C. Laugier, "Efficient GPU-based Construction of Occupancy Grids Using several Laser Range-finders," *International Journal of Autonomous Vehicles*, vol. 6, no. 1–2, pp. 48–83, 2008.

[32] A. Pouget, P. Dayan, and R. Zemel, "Information processing with population codes," *Nature Reviews Neuroscience*, vol. 1, pp. 125–132, 2000, review.

[33] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, revised second printing ed., M. B. Morgan, Ed. Morgan Kaufmann Publishers, Inc. (Elsevier), 1988.

[34] S. Treue, K. Hol, and H.-J. Rauber, "Seeing multiple directions of motion — physiology and psychophysics," *Nature Neuroscience*, vol. 3, no. 3, pp. 270–276, March 2000.

[35] R. T. Born and D. C. Bradley, "Structure and Function of Visual Area MT," *Annual Review of Neuroscience*, vol. 28, pp. 157–189, July 2005.

[36] B. Yang and J. Scheuing, "Cramer-Rao bound and optimum sensor array for source localization from time differences of arrival," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05)*, vol. 4. IEEE, 2005, pp. 961–964.

[37] J. C. Chen, K. Yao, and R. E. Hudson, "Source Localization and Beamforming," *IEEE Signal Processing Magazine*, vol. 1053, no. 5888/02, pp. 30–39, 2002.

[38] B. Loesch, P. Ebrahim, and B. Yang, "Comparison of different algorithms for acoustic source localization," in *ITG-Fachbericht-Sprachkommunikation 2010*, 2010. [Online]. Available: http://www.vde-verlag.de/proceedings-en/453300039.html

[39] T. Mukai, "Developing sensors that give intelligence to robots," *Riken Research*, vol. 1, no. 6, pp. 13–16, Jun. 2006. [Online]. Available: http://www.rikenresearch.riken.jp/eng/frontline/4428

[40] J. Laurens and J. Droulez, "Bayesian processing of vestibular information," *Biological Cybernetics*, vol. 96, no. 4, pp. 389–404, December 2007, (Published online: 5th December 2006). [Online]. Available: http://dx.doi.org/10.1007/s00422-006-0133-1

[41] ——, "Bayesian modeling of inertial self-motion perception," 2005, section 3.2 of Bayesian IBA Project Workpackage 2 deliverable D15.

[42] Y.-C. Lu, H. Christensen, and M. Cooke, "Active binaural distance estimation for dynamic sources," in *Interspeech 2007*, Antwerp, Belgium, 2007, pp. 574–577.

[43] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, November 2004.

[44] J.-Y. Bouguet, "Camera Calibration Toolbox for Matlab," 2006. [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/index.html

[45] J. Lobo, "InerVis Toolbox for Matlab," http://www.deec.uc.pt/~jlobo/InerVis_WebIndex/, 2006.

[46] J. Lobo and J. Dias, "Relative Pose Calibration Between Visual and Inertial Sensors," *International Journal of Robotics Research, Special Issue 2nd Workshop on Integration of Vision and Inertial Sensors*, vol. 26, no. 6, pp. 561–575, June 2007.

[47] R. H. S. Carpenter, "The saccadic system: a neurological microcosm," *Advances in Clinical Neuroscience and Rehabilitation*, vol. 4, pp. 6–8, 2004, review Article.

[48] A. Caspi, B. R. Beutter, and M. P. Eckstein, "The time course of visual information accrual guiding eye movement decisions," *Proceedings of the National Academy of Sciences U.S.A.*, vol. 101, no. 35, pp. 13 086–13 090, 31 August 2004.

João Filipe de Castro Cardoso Ferreira (M'12) was born in Coimbra, Portugal, in 1973. He received his B.Sc. (five-year course), M.Sc. and Ph.D. degrees in Electrical Engineering and Computers from the University of Coimbra, Portugal, in 2000, 2005 and 2011, respectively.
He has been an Invited Assistant Professor at the University of Coimbra, and a Post-Doc researcher at the Institute of Systems and Robotics (ISR) since 2011. He has also been a staff researcher at the ISR since 1999. His main research interests are human and artificial perception, robotics, Bayesian modelling and 3D scanning.
Dr. Ferreira is a member of the IEEE Robotics and Automation Society (RAS).

Jorge Nuno de Almeida e Sousa Almada Lobo (M'08) was born in Cambridge, UK, in 1971. He completed, his five-year course in Electrical Engineering, and the M.Sc. and Ph.D. degrees from the University of Coimbra, Portugal, in 1995, 2002 and 2007, respectively.
He was a junior teacher in the Computer Science Department of the Coimbra Polytechnic School, and later joined the Electrical and Computer Engineering Department of the Faculty of Science and Technology at the University of Coimbra, where he currently works as Assistant Professor. His current research is carried out at the Institute of Systems and Robotics, University of Coimbra. His current research interests focus on inertial sensor data integration in computer vision systems, Bayesian models for multimodal perception of 3D structure and motion, and real-time performance using GPUs and reconfigurable hardware.
Dr. Lobo is a member of the IEEE Robotics and Automation Society (RAS).

Pierre Bessière was born in 1958. He received the engineering degree and the Ph.D. degree in computer science from the Institut National Polytechnique de Grenoble (INPG), France, in 1981 and 1983, respectively.
He did a Post-Doctorate at SRI International (Stanford Research Institute) working on a project for National Aeronautics and Space Administration (NASA). He then worked for five years in an industrial company as the leader of different artificial intelligence projects. He came back to research in 1989. He has been a senior researcher at Centre National de la Recherche Scientifique (CNRS) since 1992. His main research concerns have been evolutionary algorithms and probabilistic reasoning for perception, inference and action. He leads the Bayesian Programming research group (Bayesian-Programming.org) on these two subjects. Fifteen PhD diplomas and numerous international publications are fruits of the activity of this group during the last 15 years. He also leads the BIBA (Bayesian Inspired Brain and Artefact) and was a Partner in the BACS (Bayesian Approach to Cognitive Systems) European project. He is a co-founder and scientific adviser of the ProBAYES Company, which develops and sells Bayesian solutions for the industry.

Miguel de Sá e Sousa Castelo-Branco received the M.D. degree from the University of Coimbra, Coimbra, in 1991, and the Ph.D. degree from the Max-Planck Institute für Hirnforschung, Frankfurt, and the University of Coimbra, in 1998.
He 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, which is a part of the European Network Evi-Genoret. He is also involved in the Portuguese National Functional Brain Imaging Network. He has made contributions in the fields of ophthalmology, neurology, visual neuroscience, human psychophysics, functional brain imaging and human and animal neurophysiology.

Jorge Manuel Miranda Dias (M'96-SM'10) received his Ph.D. in Electrical Engineering with specialisation in Control and Instrumentation from the University of Coimbra, Portugal, in 1994.
He holds his research activities at the Institute of Systems and Robotics (Instituto de Sistemas e Robótica), University of Coimbra, and also at the Khalifa University of Science, Technology and Research (KUSTAR), Abu Dhabi, UAE. His current research areas are computer vision and robotics, with activities and contributions in these fields since 1984. He has been the main researcher in several projects financed by the European Commission (Framework Programmes 6 and 7) and by the Portuguese Foundation for Science and Technology (FCT).
Dr. Dias 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".