Bayesian Models for Multimodal Perception of 3D Structure and Motion

J.F. Ferreira, Pierre Bessière, Kamel Mekhnacha, J. Lobo, J. Dias, Christian Laugier

Proceedings of the 2008 International Conference on Cognitive Systems
University of Karlsruhe, Karlsruhe, Germany, April 2-4, 2008

Bayesian Models for Multimodal Perception of 3D Structure and Motion

João Filipe Ferreira*, Pierre Bessière†, Kamel Mekhnacha‡, Jorge Lobo*, Jorge Dias* and Christian Laugier§
*ISR — Institute of Systems and Robotics, FCT-University of Coimbra, Coimbra, Portugal
†CNRS-Grenoble, France
‡Probayes, Grenoble, France
§INRIA Rhône-Alpes, Grenoble, France

***Abstract*— In this text we will formalise a novel solution, the Bayesian Volumetric Map (BVM), as a framework for a metric, short-term, egocentric spatial memory for multimodal perception of 3D structure and motion. This solution will enable the implementation of top-down mechanisms of attention guidance of perception towards areas of high entropy/uncertainty, so as to promote active exploration of the environment by the robotic perceptual system. In the process, we will to try address the inherent challenges of visual, auditory and vestibular sensor fusion through the BVM. In fact, it is our belief that perceptual systems are unable to yield truly useful descriptions of their environment without resorting to a temporal series of sensory fusion processed on a short-term memory such as the BVM.**

I. 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 perception models.

In this text we will formalise a novel framework for a metric, short-term, egocentric spatial memory that will enable the implementation of top-down mechanisms of attention guidance of perception towards areas of high uncertainty, so as to promote active exploration of the environment by the robotic perceptual system.

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 Fig. 1.

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.

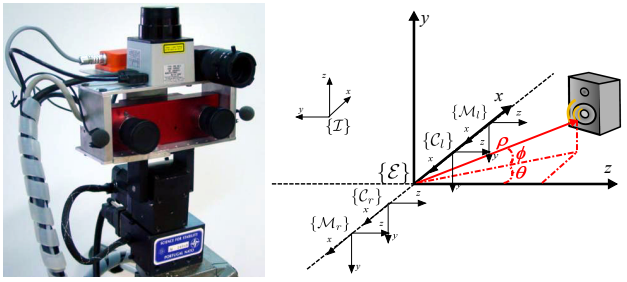

Fig. 1. View of the current version of the Integrated Multimodal Perception Experimental Platform (IMPEP), on the left. On the right, the IMPEP perceptual geometry is shown: $\{\mathcal{E}\}$ is the main reference frame for the IMPEP robotic head, representing the egocentric coordinate system;$\{\mathcal{C}_{l,r}\}$ 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.

II. Background and Definitions

The perceptual brain is known to be divided into two separate, albeit interdependent, pathways of sensory processing: the *ventral pathway*, popularly denominated as the "what" pathway, which is concerned primarily with perceptual classification and recognition tasks, and the *dorsal pathway*, popularly known as the "where" pathway, which is dedicated to fast processing of sensory information, with the sole purpose of yielding spatial representations (e.g. positioning, structure and motion), whatever the nature of the entity it is analysing. Given the perceptual problem exposed earlier on, the latter is of particular interest to our work. These spatial representations are believed to be *metric* and *egocentric* in lower-level areas of the dorsal pathway so as to promote fast and accurate interaction with the surrounding environment.

Given these facts, 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 allows for the *representation of dynamical spatial occupation* of the environment, avoiding any need for any assumptions on the nature of those objects, or in other words, for data association. Data association is thus effectively postponed to higher-level processing.

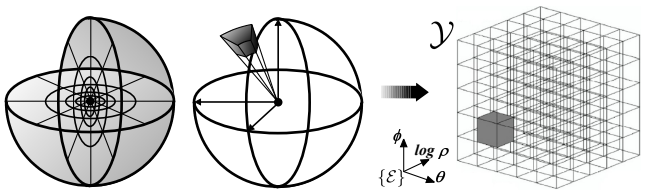

Fig. 2. Egocentric, log-spherical configuration of the Bayesian Volumetric Maps.

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.

In our specific application domain, common occupancy grid configurations which assume regularly partitioned euclidean space to build the cell lattice are not appropriate. Hence, we chose a *log-spherical* coordinate system spatial configuration (see Fig. 2) for our *Bayesian Volumetric Maps* (BVM), thus promoting an egocentric trait and yielding more precision for objects closer to the observer, which seems to agree 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.

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:

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

where constant values for log-distance base b, and angular ranges $\Delta\theta = \theta_{\text{max}} - \theta_{\text{min}}$ and $\Delta\phi = \phi_{\text{max}} - \phi_{\text{min}}$, 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 $C = (\log_b \rho_{\text{max}}, \theta_{\text{max}}, \phi_{\text{max}})$.

More recently, Coué *et al.* [3] and Tay *et al.* [4] expanded on the occupancy grid by explicitly introducing *Bayesian filtering*. A two-step mechanism estimates, at each time step, the state of the occupancy grid by combining a prediction step (history) and an estimation step (incorporating new measurements). This approach is derived from the Bayesian filtering approach [5], and is thus named the *Bayesian Occupancy Filter* (BOF).

To compute the probability distributions for the current states of each cell, the *Bayesian Program* (BP) formalism, as first defined by Lebeltel [6], will be used throughout this text.

III. Bayesian Volumetric Maps for Multimodal Perception

A. Sensor fusion advantages and challenges

The use of more than one sensor promotes a robustness increase on the observation and characterisation of a physical phenomenon. In fact, using different types of sensors allows for the dilution of each sensor's individual weaknesses through the use of the strengths of the remainder.

There is evidence that humans fuse perceptual cue information following mainly two general strategies [7]: *combination*, that expresses interactions between sensory signals that are not redundant, and *integration*, that expresses interactions between sensory signals that are redundant. Combination has the purpose of maximising information coming from different cues, whilst the goal of integration is to minimise variance in the sensory estimate to increase its reliability. For several estimates resulting from combination to be integrated into a single estimate, they must be in the same units and referred to the same coordinate system, and hence must undergo a process called *promotion* [7].

We will try to explicitly or implicitly address each of the challenges of sensor fusion as described in [7] using the BVM, 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 a short-term memory such as the BVM. We propose to use vestibular sensing as ancillary information to promote visual and auditory sensing to satisfy the requirements for integration, enumerated above.

B. Using Bayesian filtering for visuoauditory integration

The Bayesian Program presented in Fig. 3 is based on the solution presented by Tay *et al.* [4], 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. 3. This inference effectively leads to the Bayesian filtering formulation as used in the BOF grids — see Fig. 4. 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 $P(O_C\, V_C|Z_1 \cdots Z_S\, C)$. The result

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 possible cases $\in \mathcal{V} \equiv \{v_1, \cdots, v_n\}$.

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:

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

Fig. 3. Bayesian Program for the estimation of Bayesian Volumetric Map current cell state.

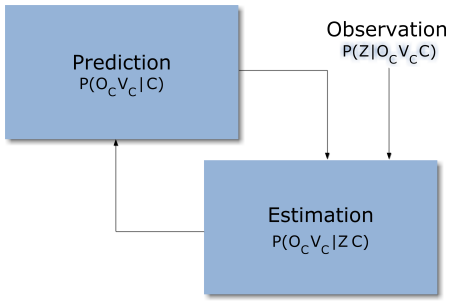

Fig. 4. Bayesian filtering for the estimation of occupancy and local motion distributions in the BVM. The schematic considers only a single measurement for simpler reading, with no loss of generality.

from the Bayesian filter estimation will then be used for the prediction step in the next iteration.

C. 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.

D. 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 [8], the BVM may be used safely to integrate auditory and vision measurements as soon they become available; preliminary tests using the BVM update model show that this, in fact, promotes an effect similar to the well-known temporal ventriloquism, given the inherent auditory measurement frequency as opposed to vision. Spatial ventriloquism, on the other hand, is implicitly ensured due to the inherent properties of the Bayesian formulation of visuoauditory integration (i.e. modality reliability expressed in terms of uncertainty). Promotion through vestibular sensing is also not a problem, since inertial readings are available at a much faster rate than visuoauditory perception.

IV. Sensor Models

A. Vision sensor model

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 [9]. The stereovision algorithm used for visual depth sensing is an adaptation of the fast and simple coherence detection approach by Henkel [10], 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 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 [11]; 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.

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.* [12].

In the spirit of Bayesian programming, we again start by defining the relevant variables:

- $C \equiv (\log_b \rho_{\max}, \theta_{\max}, \phi_{\max})$, O_C and Z have the same meaning as before. However, once a projection line (θ,ϕ), with $\theta_{\min} \leq \theta \leq \theta_{\max} \wedge \phi_{\min} \leq \phi \leq \phi_{\max}$, is established for a sensor, only $\log_b \rho_{\max}$ varies throughout the respective line-of-sight, thus effectively indexing each cell. Therefore, by abuse of notation and in order to simplify references to cells in the line-of-sight, these will be referred to using the abstraction $C \in \mathbb{N}, 1 \leq C \leq N$, where $N = \log_b(\rho_{Max} - \rho_{Min})$ denotes the total number of cells in the line-of-sight.
- $G_C \in \mathcal{G}_C \equiv \mathcal{O}^{N-1}$ represents the state of all cells in the line-of-sight except for C. Each g_C is, thus, an $(N-1)$-tuple of the form $([O_1 = o_1], \cdots, [O_{c-1} = o_{c-1}], [O_{c+1} = o_{c+1}], \cdots, [O_N = o_N])$ given a specific cell $[C = c]$.

The following expression gives the decomposition of the joint distribution of the relevant variables according to Bayes' rule and dependency assumptions:

$$P(Z\,C\,O_C\,G_C) = P(C)P(O_C|C)P(G_C|O_C\,C)P(Z|G_C\,O_C\,C) \tag{3}$$

The parametric form and semantics of each component of the joint decomposition are then as follows:

- $P(C)$ and $P(O_C|C)$ represent *a priori* information on the environment. Note that $P(C)P(O_C|C)$ is, in fact, formally equivalent to $P(A_C)P(V_C|A)P(C|V_C\,A_C)P(O^{-1}|A)P(O|O^{-1})$ when considering scene dynamics. The probability of a cell being empty is $P_{\text{Empty}} = P([O_C = 0]|C)$.
- $P(G_C|O_C\,C) \equiv P(G_C|C)$ represents the probability that, knowing a state of a cell, the whole line-of-sight is in a particular state [12].
- $P(Z|G_C\,O_C\,C)$ is sensor-dependent but, in any case, for all $(O_C, G_C) \in \mathcal{O} \times \mathcal{G}_C$, the probability distribution over Z depends only on the *first occupied cell*. Knowing the position of the first occupied cell in the projection line, which will be denoted as $[C = k]$, $P(Z|G_C\,O_C\,[C = k])$ gives the probability of a measurement if $[C = k]$ would be the only occupied cell in the line-of-sight. This particular distribution over Z is called the *elementary sensor model*, denoted by $P_k(Z)$.

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

$$P_k(Z) = L_k(Z, \mu_\rho(k), \sigma_\rho(k)),\ \begin{cases} \mu_\rho(k) & = \hat{\rho}(\hat{\delta}) \\ \sigma_\rho(k) & = \frac{1}{\lambda}\sigma_{min} \end{cases} \tag{4}$$

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.

Equation (4) 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 [13], [14], [1], [9].

However, the fact that depth sensors always yield positive readings may be contradicted by the circumstance that normal distributions assign non-zero probabilities to negative depth values; even worse, close to the origin ($Z = 0$) this distribution assigns a *high* probability to negative depth values! With this purpose, we have adapted Yguel *et al.*'s Gaussian elementary sensor model so as to additionally perform the transformation to distance log-space, as follows

$$P_k([Z = z]) = \begin{cases} \int_{]-\infty;1]} \mathcal{N}(\mu(k-0.5), \sigma(\sigma_\rho))(u)du, & z \in [0;1] \\ \int_{\lfloor z \rfloor}^{\lfloor z \rfloor + 1} \mathcal{N}(\mu(k-0.5), \sigma(\sigma_\rho))(u)du, & z \in]1;N] \\ \int_{]N;+\infty]} \mathcal{N}(\mu(k-0.5), \sigma(\sigma_\rho))(u)du, & z = \text{"No Detection"} \end{cases} \tag{5}$$

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 $]k-1;k]$.

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 given in (3); 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.

To correctly formalise the Bayesian inference process, a formal auxiliary definition with respective properties follow.

Definition 1: $A_c^k \in \mathcal{G}_C$ is the set of all tuples for which the first occupied cell is $[C=k]$. Formally, it denotes tuples such as $(o_1,\cdots,o_{c-1},o_{c+1},\cdots,o_N) \in \{0,1\}^{N-1}$, yielding $[O_i=0]\wedge[O_k=1], \forall i<k$.

Property 1.1: $\forall(i,j), i\neq j, A_c^i \bigcap A_c^j = \emptyset$

Property 1.2: $\bigcup A_c^k = \mathcal{G}_c \setminus \mathcal{G}_\emptyset$, with
$\mathcal{G}_\emptyset = \{(o_p)_p | \forall p \in \mathbb{N}\setminus\{c\},\, 1\leq p\leq N, [O_p=0]\}$

Property 1.3: If $k<c$ there are k determined cells: the $k-1$ first cells, $(o_1,\cdots,o_{k-1})$, which are empty, and the kth, (o_k), which is occupied. *Then*, $P(A_c^k) = P_{\text{Empty}}^{k-1}(1-P_{\text{Empty}})$.

Property 1.4: If $k>c$ there are $k-1$ determined cells: the $k-2$ first cells, $(o_1,\cdots,o_{c-1},o_{c+1},\cdots,o_{k-1})$, which are empty, and the $(k-1)$th, (o_k), which is occupied. *Then*, $P(A_c^k) = P_{\text{Empty}}^{k-2}(1-P_{\text{Empty}})$.

It now becomes possible to determine $P(Z|O_C\,C)$ in order to express the desired joint distribution $P(Z\,O_C\,C)$. This process leads to four distinct possible cases, that will be described next.

In the case of detection given an occupied cell $[C=c]$, the sensor measurement can only be due to the occupancy of this cell or a cell before it in terms of visibility.

Thus [12],

$$\begin{aligned}
&\forall Z \neq \text{“No Detection”},\\
&P(Z|[O_C=1]\,C) =\\
&\quad= \sum_{g_c\in\mathcal{G}_C} P([G_c=g_c])P(Z|[O_C=1]\,[G_c=g_c]\,C)\\
&\quad= \sum_{k=1}^{c-1} P(A_c^k)P_k(Z) + (1-\sum_{k=1}^{c-1} P(A_c^k))P_c(Z)\\
&\quad= \sum_{k=1}^{c-1} P_{\text{Empty}}^{k-1}(1-P_{\text{Empty}})P_k(Z) + P_{\text{Empty}}^{k-1}P_c(Z)
\end{aligned} \tag{6}$$

Equation (6) has two terms: the left term that represents the case where $[C=c]$ is occupied and the right term that comes from the aggregation of all the remaining probabilities around the last possible cell that might produce a detection: $[C=c]$ itself. The “No Detection” case ensures that the distribution is normalised.

In the case of no detection given an occupied cell $[C=c]$, which would correspond most probably to the effects of occlusion from earlier cells,

$$\begin{aligned}
&Z = \text{“No Detection”},\\
&P(Z|[O_C=1]\,C) =\\
&\qquad= 1 - \sum_{r\neq\text{“No Det.”}} P([Z=r]|[O_C=1]\,C)
\end{aligned} \tag{7}$$

In the case of a measurement from detection knowing that $[C=c]$ is empty, where a erroneous detection is yielded by the sensor (the so-called *false alarm*),

$$\begin{aligned}
&\forall Z \neq \text{“No Detection”},\\
&P(Z|[O_C=0]\,C) =\\
&\quad= \sum_{g_c\in\mathcal{G}_C} P([G_c=g_c])P(Z|[O_C=0]\,[G_c=g_c]\,C)\\
&\quad= \sum_{k=1,k\neq c}^{N} P(A_c^k)P_k(Z) + P(\mathcal{G}_\emptyset)\delta_{Z=\text{“No Detection”}}\\
&\quad= \sum_{k=1}^{c-1} P_{\text{Empty}}^{k-1}(1-P_{\text{Empty}})P_k(Z) +\\
&\quad+ \sum_{k=c+1}^{N} P_{\text{Empty}}^{k-2}(1-P_{\text{Empty}})P_k(Z) + P_{\text{Empty}}^{N-1}\delta_{Z=\text{“No Det.”}}
\end{aligned} \tag{8}$$

There are three terms in the empty cell, from left to right, corresponding respectively to before the detection, after the detection and no detection at all. Again, the “No Detection” case ensures that the distribution is normalised.

In the case of no detection knowing that $[C=c]$ is empty, which will either be due to a miss-detection or a completely empty line-of-sight corresponding to $\mathcal{G}_\emptyset$,

$$\begin{aligned}
&Z = \text{“No Detection”},\\
&P(Z|[O_C=0]\,C) =\\
&\quad= 1 - (\sum_{r}^{N} P([Z=r]|[O_C=0]\,C)) + P_{\text{Empty}}^{N-1}\delta_{Z=\text{“No Det.”}}
\end{aligned} \tag{9}$$

B. Audition sensor model

The audition sensor model used as a source of observations for BVM cell state updates is fully described in [15].

C. Vestibular sensor model

To process the inertial data, we adapted the Bayesian model proposed by Laurens and Droulez [16] 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. Since we only consider the simplest case, where sensors may only rotate around the egocentric origin, the angular rotation measurements may be safely assumed to be independent; linear acceleration might have a centripetal component that depends on the distance to the origin, but since the model is only detecting gravity, only a sustained rotation and a

significant distance to the origin would produce an error in the angular position, like when a test pilot is in a centrifuge.

In this model, X, Y and Z refer to the three axes of the robotic vision head in egocentric coordinates. The orientation of the system in space is encoded using a rotation matrix $\boldsymbol{\Theta}$. Angular velocity of the head is encoded using the yaw y, pitch p and roll r conventions. Yaw rotations are rotations around the Z axis; pitch around the Y axis and roll around X. When a rotation consists of a combination of yaw, pitch and roll rotation, the three rotations are applied successively and in this order. The rotation update is given by

$$\boldsymbol{\Theta}^{t+\delta t} = \boldsymbol{\Theta}^{t}.\boldsymbol{R}(\delta y, \delta p, \delta r) \tag{10}$$

where $\boldsymbol{R}(y,p,r) =$

$$\begin{bmatrix} \mathrm{c}(y).\mathrm{c}(p) & \mathrm{c}(y).\mathrm{s}(p).\mathrm{s}(r) - \mathrm{s}(y).\mathrm{c}(r) & \mathrm{c}(y).\mathrm{s}(p).\mathrm{c}(r) + \mathrm{s}(y).\mathrm{s}(r) \\ \mathrm{s}(y).\mathrm{c}(p) & \mathrm{c}(y).\mathrm{c}(r) + \mathrm{s}(y).\mathrm{s}(p).\mathrm{s}(r) & -\mathrm{c}(y).\mathrm{s}(r) + \mathrm{s}(y).\mathrm{s}(p).\mathrm{c}(r) \\ -\mathrm{s}(p) & \mathrm{c}(p).\mathrm{s}(r) & \mathrm{c}(p).\mathrm{c}(r) \end{bmatrix}$$

and the instantaneous angular velocity is defined as:

$$\boldsymbol{\Omega} = \begin{pmatrix} \delta y/\delta t \\ \delta p/\delta t \\ \delta r/\delta t \end{pmatrix}$$

The calibrated inertial sensors in the IMU provide direct egocentric measurements of body angular velocity and linear acceleration (including gravity $\boldsymbol{G}$). Given the motion of the system, we can define the probability distribution of the sensory inputs. The gyros will measure $\boldsymbol{\Omega}^t$ with added Gaussian noise, i.e. $\boldsymbol{\Phi}^t = \boldsymbol{\Omega}^t + \eta_\Phi^t$, where η_Φ^t is a three-dimensional vector, the elements of which follow independent Gaussian distributions with mean 0 and standard deviation σ_Φ. The accelerometers will measure the gravito-inertial acceleration $\boldsymbol{F}$ with added Gaussian noise, i.e. $\boldsymbol{\Upsilon}^t = \boldsymbol{F}^t + \eta_\Upsilon^t$, where η_Υ^t is a three-dimensional vector, the elements of which follow independent Gaussian distributions with mean 0 and standard deviation σ_Υ. $\boldsymbol{F}$ is the resultant acceleration due to linear acceleration and gravity. Given the geocentric body linear acceleration $\boldsymbol{A}$ and the system orientation $\boldsymbol{\Theta}$, we can compute $\boldsymbol{F}$. Transforming to the egocentric frame of reference we have

$$\boldsymbol{F} = \boldsymbol{\Theta}^{-1}.(\boldsymbol{G} - \boldsymbol{A}) \tag{11}$$

The sensor data at time t is therefore defined by $(\boldsymbol{\Phi}^t, \boldsymbol{\Upsilon}^t)$, and the state of our system at time t by $(\boldsymbol{\Theta}^t, \boldsymbol{\Omega}^t, \boldsymbol{A}^t)$. Estimates for spherical angles $(\theta_{\text{inertial}}, \phi_{\text{inertial}})$ are then easily derived from the pitch-roll-yaw configuration of $\boldsymbol{\Theta}$.

As suggested in [16], even in the absence of any sensory information, motion estimates for which the rotational velocity and acceleration are low are more probable. This can be described in a simple way using a Gaussian distribution. Having

$$\mathcal{N}(x, \mu, \sigma) = \frac{e^{-(x-\mu)^2/(2.\sigma^2)}}{\sqrt{2.\pi.\sigma^2}}$$

the probability distribution for acceleration is given by $P(\boldsymbol{A}^t) \propto \mathcal{N}(|\boldsymbol{A}^t|, 0, \sigma_A)$; similarly for angular velocity $\boldsymbol{\Omega}$ we have $P(\boldsymbol{\Omega}^t) \propto \mathcal{N}(|\boldsymbol{\Omega}^t|, 0, \sigma_\Omega)$.

V. Conclusions

We have formalised herewith a novel solution, the Bayesian Volumetric Map, a framework for a metric, short-term, egocentric spatial memory for multimodal perception of 3D structure and motion. This solution allows: the estimation of 3D structure and local motion states through perceptual fusion, involving vision, audition and inertial sensing, effectively postponing data association to higher-level perceptual processing; the implementation of top-down mechanisms of attention guidance of perception towards areas of high entropy/uncertainty, so as to promote active exploration of the environment by the robotic perceptual system.

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

References

[1] 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.

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

[3] C. Coué, C. Pradalier, C. Laugier, T. Fraichard, and P. Bessière, "Bayesian occupancy filtering for multitarget tracking: an automotive application," *Int. Journal of Robotics Research*, vol. 25, no. 1, pp. 19–30, 2006.

[4] 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," 2007, *International Journal of Autonomous Vehicles*.

[5] A. H. Jazwinsky, *Stochastic Processes and Filtering Theory*. New York: Academic Press, 1970, iSBN 0-12381-5509.

[6] O. Lebeltel, "Programmation Bayésienne des Robots," Ph.D. dissertation, Institut National Polytechnique de Grenoble, Grenoble, France, September 1999.

[7] 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.

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

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

[10] R. Henkel, "A Simple and Fast Neural Network Approach to Stereovision," in *Proceedings of the Conference on Neural Information Processing Systems — NIPS'97*, M. Jordan, M. Kearns, and S. Solla, Eds. Denver: MIT Press, Cambridge, 1998, pp. 808–814.

[11] 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.

[12] M. Yguel, O. Aycard, and C. Laugier, "Efficient GPU-based Construction of Occupancy Grids Using several Laser Range-finders," 2007, *International Journal of Autonomous Vehicles*.

[13] 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.

[14] 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.

[15] 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.

[16] J. Laurens and J. Droulez, "Bayesian processing of vestibular information," *Biological Cybernetics*, December 2006, (Published online: 5th December 2006). [Online]. Available: http://dx.doi.org/10.1007/s00422-006-0133-1