UNIVERSIDADE DE COIMBRA

FACULDADE DE CIÊNCIAS E TECNOLOGIA

Departamento de Engenharia Electrotécnica e Computadores

PH. D. THESIS

Bayesian Cognitive Models for 3D Structure and Motion Multimodal Perception

João Filipe Ferreira

Coimbra, 2010

UNIVERSIDADE DE COIMBRA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
Departamento de Engenharia Electrotécnica e Computadores

Ph. D. Thesis

Bayesian Cognitive Models for 3D Structure and Motion Multimodal Perception

João Filipe Ferreira

Coimbra, 2010

Dissertation submitted to the Department of Electrical Engineering and Computer Science of the University of Coimbra, partially satisfying the requirements for a Doctorate degree.

Research work developed under the supervision of Professor Jorge Manuel Miranda Dias — Associate Professor at the Department of Electrical Engineering and Computer Science of the University of Coimbra — and co-supervision of Professor Miguel de Sá e Sousa Castelo-Branco — Assistant Professor at the Faculty of Medicine of the University of Coimbra.

This work is dedicated to the better part of me:
my wife, Anita, and my children, Luísa and Miguel.

Human thinking and probability theory:

Probability theory is nothing but common sense reduced to calculation.

Laplace, 1819.

Dreaming the future of Artificial Intelligence and Robotics against all scepticism:

You insist that there is something a machine cannot do. If you tell me ***precisely*** *what it is that a machine cannot do, then I can always make a machine which will do just that!*

J. von Neumann in a famous 1948 talk on computers as a reply to the canonical question from the audience: “But a mere machine can’t really *think*, can it?”

The cross-disciplinary conundrum:

What am I supposed to publish?

L.J. Savage (1962, *The Foundations of Statistical Inference, a Discussion*, Methuen, London) asked this question to express his bemusement at the fact that, no matter what topic he chose to discuss, and no matter what style of writing he chose to adopt, he was sure to be criticised for not making a different choice.
He was not alone [Jaynes 2003].

Acknowledgements

First and foremost, I am grateful to my supervisors, Prof. Jorge Dias, for enduring with my sometimes stubborn and unsubtle demeanour since my Master's work, and Prof. Castelo-Branco, for taking on the challenge of training this robotic-oriented student in the "mysterious ways" of reasoning in Neuroscience terms. I have become an infinitely better researcher and human being due to their wisdom and innovative and energetic spirit, and also their willingness in taking me on this amazing, albeit sometimes bumpy ride through the wonderful world of bioinspired multimodal perception and cognition.

Just as I was starting to despair while thinking how on earth I would get a probabilistic solution for my problem, I was offered the incredible opportunity of joining INRIA's e-Motion team in Grenoble through the European project BACS and share an office with Prof. Pierre Bessière, for what turned out to be what I now consider the two most productive months of my PhD research work. Pierre was, in fact, what one might call the "unofficial co-supervisor" of this thesis.

From beginning to end, I was always able to rely upon excellent team work with Prof. Jorge Lobo, who was always kind enough to put his scientific and engineering spirit at the service of this work, and to whom I will always be in debt for his enduring friendship and counselling.

In the course of this ride, I have also had the opportunity to work with many incredibly talented people. I would particularly like to thank, at the Institute of Systems and Robotics (ISR/FCT-UC), Cátia Pinho, former collaborator, for her work on the binaural calibration procedure and the BVM viewer software in MATLAB, Luís Santos, for the realtime BVM viewer software in C++/OpenGL, Pedro Trindade for putting his friendship at my service through his always kind, reassuring and wise suggestions, while also helping me by implementing an upgraded Blender-based version of the BVM viewer software, João Quintas for his inestimable help with the concluding experimental work, and finally Alex Malhão and Hugo Faria for their help with the robotic head; at the Institute of Biomedical Research in Light and Image (IBILI/UC), I would like to thank Britta Gräwe, my "partner in crime" in psychophysics and human studies experiments, and also Gil Cunha and João Castelhano for their help in the final and pilot experiments; in Grenoble, I would like to thank Manuel Yguel (INRIA Rhône-Alpes) for his invaluable help with the vision sensor model, Cristopher Tay (INRIA Rhône-Alpes) and Kamel Mekhnacha (Probayes) for their help with the implementation of BOF framework, and finally everybody at Probayes for their help with the ProBT implementations; in Switzerland, Ricardo Chavarriaga and Sara Gonzalez for their help on developing the hierarchical model, both by keeping me grounded to my bioinspired roots and by inciting me to follow the potential for future developments of this work. I would also like to thank all the previously mentioned people and all my other colleagues at the Mobile Robotics Lab at the ISR and at the IBILI, and also from the BACS project, for their numerous and invaluable suggestions and voluntary participation on most of the demos and experimental work. There is simply too many of you to mention herewith.

Last but definitely not least, I am immensely indebted to my parents, my sister and all my friends, for making me what I am today and for **always** having been there for me; however, I owe everything to my wife Anita, and my daughter and son Maria Luísa and Miguel Francisco, my main reasons for carrying on. In particular for you, I hope I made you proud and that you felt my ever-present love for you despite my frequent absences from home (even when at home!).

All research work leading to this thesis was supported by a FCT scholarship (SRFH/BD/24628/2005) granted since January 2006 until January 2010, and also by the European project BACS — Bayesian Approach to Cognitive Systems (contract no. FP6-IST-027140).

Abstract

Humans use various sensory cues to extract crucial information from the environment. With a view of having robots as human companions, we are motivated towards helping to develop a knowledge representation system along the lines of what we know about us. While recent research has shown interesting results, we are still far from having concepts and algorithms that interpret space, coping with the complexity of the environment.

By understanding how animals (humans) navigate and build their own spatial representation, the observed phenomena can be applied in robotics. In order to have a robust and reliable framework for navigation (i.e. in order to move within an environment, manipulate objects in it, avoid undesirable mishaps — e.g. collisions — etc.) space representation, localisation, mapping and perception are all needed.

The goal of this work was to research Bayesian models to deal with fusion, multimodality, conflicts, and ambiguities in perception, while simultaneously drawing inspiration upon human perceptual processes and respective behaviours.

We will present a Bayesian framework for active multimodal perception of 3D structure and motion which, while not strictly neuromimetic, finds its roots in the role of the dorsal perceptual pathway of the human brain. Its composing models build upon a common egocentric spatial configuration that is naturally fitting for the integration of readings from multiple sensors using a Bayesian approach. At its most basic level, these models present efficient and robust probabilistic solutions for cyclopean geometry-based stereovision and auditory perception based only on binaural cues, defined using a consistent formalisation that allows their use as building blocks for the multimodal sensor fusion framework, both explicitly or implicitly addressing the most important challenges of sensor fusion, for vision, audition and proprioception (including vestibular sensing). Parallely, baseline research on human multimodal motion perception presented in this text provides the support for future work in new sensor models for the framework. This framework is then extended in a hierarchical fashion by incrementally implementing active perception strategies, such as active exploration

based on entropy of the perceptual map that constitutes the basis of the framework and sensory saliency-based behaviours.

The computational models described herewith support a real-time robotic implementation of multimodal active perception to be used in real-world applications, such as human-machine interaction or mobile robot navigation.

With this work, we also hope to be able to address questions such as: Where are the limits on optimal sensory integration behaviour? What are the temporal aspects of sensory integration? How do we solve the “correspondence problem” for sensory integration? How to answer the combination versus integration debate? How to answer the switching versus weighing controversy? What are the limits of crossmodal plasticity?

Sumário

Para extrair informação crucial do meio circundante, os seres humanos recorrem a pistas provenientes de múltiplas fontes sensoriais. Tendo como objectivo a utilização de robôs como companheiros, somos motivados no sentido de desenvolver um sistema de representação de conhecimento inspirado no que sabemos sobre o Homem. Apesar dos últimos desenvolvimentos, sem dúvida importantes, resultantes da investigação mais recente nesta área, estamos ainda longe de ter chegado a conceitos e algoritmos que interpretem o espaço e que sejam simultaneamente capazes de lidar com a complexidade do meio ambiente.

Através da compreensão de como os animais (mais concretamente, o Homem) navegam e constroem as suas próprias representações do espaço circundante, os fenómenos observados podem ser aplicados em robótica. De forma a obter-se um sistema robusto e fiável para navegação (isto é, para coordenar o movimento do robô no seu ambiente, manipular objectos nesse ambiente, evitar contratempos indesejados, como colisões, etc.), representação e localização espacial, mapeamento e percepção são todos essenciais.

O objectivo deste trabalho consistiu na investigação de modelos probabilísticos baseados na regra de Bayes, capazes de lidar com fusão multissensorial, e conflitos e ambiguidades perceptuais, inspirados na percepção humana e comportamentos respectivos.

Iremos apresentar um sistema probabilístico para percepção multimodal activa de estrutura e movimento tridimensionais que, apesar de não ser neuromimética no sentido estrito, encontra as suas raízes no papel desempenhado pelo sistema perceptual dorsal do cérebro humano. Os seus modelos constituintes baseiam-se numa representação espacial egocêntrica comum, representação esta que se adequa de forma natural à integração de leituras provenientes de diferentes sensores, usando uma abordagem probabilística baseada na regra de Bayes. No seu nível mais básico, este modelos constituem soluções eficientes e robustas para estereovisão e percepção auditiva baseada unicamente em grandezas binaurais, e são definidos usando um formalismo matemático

consistente, que permite o seu uso como blocos de construção para um sistema de fusão multissensorial; por sua vez, este sistema pretende ajudar a resolver desafios importantes em termos de fusão de sensações visuais, auditivas e resultantes de propriocepção (incluindo percepção vestibular). Paralelamente, a investigação base em percepção multissensorial humana de movimento apresentada nesta dissertação suportará trabalho futuro em novos modelos de sensor. O sistema multissensorial é depois complementado de uma forma hierárquica através da modelação incremental de estratégias de percepção activa, tal como a exploração activa baseada na entropia do mapa perceptual que constitui a base do sistema e também os comportamentos resultantes de saliência sensorial. Os modelos computacionais descritos neste texto suportam um sistema robótico com capacidade de funcionamento em tempo-real, a ser utilizado em aplicações práticas, tal como na interacção homem-máquina ou na navegação de robôs móveis.

Com este trabalho, queremos também tentar responder a perguntas como: Quais os limites para a integração multissensorial óptima? Quais os aspectos temporais dessa integração? Como resolver o "problema da correspondência"? Como responder à polémica "combinação *versus* integração"? Como responder à controvérsia "comutação *versus* pesagem"? Quais os limites da plasticidade intermodal?

Contents

List of Figures

Chapter 1

Introduction

1.1 General Motivations

Consider the following scenario (Fig. 1.1) — a moving 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, motion trajectory and velocity of all entities in the scene, while taking into account the ambiguities and conflicts inherent to the perceptual process?

Both humans and robots alike operate in a world of sensory uncertainty. Robotics researchers are used to having to deal with perceptual error propagation from sensor accuracy and precision ratings, discretisation due to analogue-to-digital transformations, approximation truncations, and round-off effects from numeric representation limitations (i.e., finite number of digits used by digital memories and processing units). When considering biological perception systems, introspection fools us into thinking that perception is deterministic and certain; however, many factors contribute to limiting the reliability of information about the world taken using biological sensors — ambiguity due to physical constraints (e.g. the mapping of 3D objects into 2D images, or the "aperture problem" in local motion detection), neural noise introduced in the early stages of sensory coding, and structural constraints on neural representations and computations (e.g. the density of receptors in the retina — see, for example, Silva, Maia-Lopes, Mateus, Guerreiro, Sampaio, Faria, and Castelo-Branco [2008] — the biological counterpart of discretisation) [Knill and Pouget 2004].

Indeed, any model of a real phenomenon is incomplete — hidden variables, not taken into account in the model, influence it. The effect of these hidden variables is that the model and the phenomenon never behave exactly alike. *Uncertainty* is the direct and unavoidable consequence of *incompleteness*. No model may foresee exactly

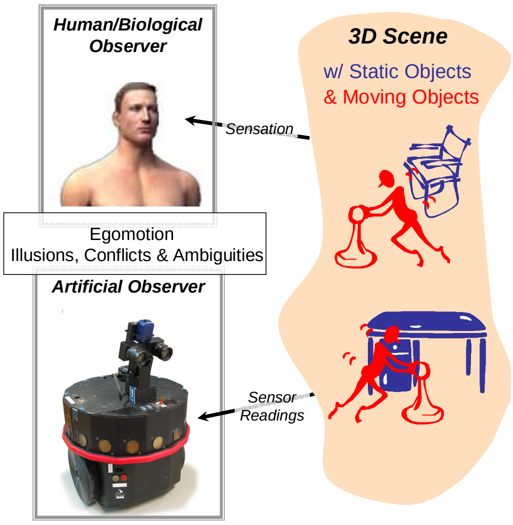


Figure 1.1: Setting for the perception of 3D structure, ego- and independent motion (human observer image courtesy of 3DScience.com).

future observations, as these observations are biased by the hidden variables. No model may either predict exactly the consequences of its decisions [Colas, Diard, and Bessière 2010].

Within the various currents attempting to solve this problem, including the so-called logicists (traditional logic), calculists (fuzzy logic, Dempster-Shafer calculus, etc.) and probabilists as defined by Pearl [1988], the often partially improperly called "Bayesian approach" proposes probability theory as an alternative to symbolic logic (i.e. Boolean logic) for rational reasoning in presence of incompleteness and uncertainty [Jaynes 2003, Colas et al. 2010].

This approach deals with incompleteness and uncertainty with a two-step process: *learning* and *inference* [Colas et al. 2010]. Learning transforms the irreducible incompleteness into quantified uncertainty (i.e. probability distributions). These distributions result from both the preliminary knowledge of the reasoning subject and the experimental data coming from the observation of the phenomenon [Colas et al. 2010, Pearl 1988]. Inference is performed with the probability distributions obtained by the first step [Colas et al. 2010]. To do so, we only require the two basic rules of Bayesian inference (Bayes theorem and the normalisation rule).

On the other hand, *ambiguity* occurs when there is a possibility to be interpreted in multiple ways. Often, an ambiguity arises in the case of an ill-posed and inverse problem [Colas et al. 2010].

Sensation is commonly defined as the effect of some phenomenon on the senses.

Perception, on the other hand, is recovering information on the phenomenon, given the sensation. Indeed, it is often easy to predict what are the sensations corresponding to a particular phenomenon. In this case, the direct function yields the sensation given the phenomenon, whereas perception is the inverse problem of extracting the phenomenon given the sensation [Poggio 1984, Yuille and Bülthoff 1996, Pizlo 2001]. In the Bayesian framework, an inverse problem such as this is addressed using the symmetry of Bayes' rule [Colas et al. 2010]. Moreover, *perception is very often an ill-posed problem* — this is the case for most multistable percepts, where a given stimulus can be perceived consistently in different ways [Colas et al. 2010, Castelo-Branco et al. 2002, Kozak and Castelo-Branco 2008].

Perception has thus, unsurprisingly, as of recently been regarded as a computational process of unconscious, probabilistic inference (although in the nineteenth century Herman von Helmholtz had already suggested this to be true for visual perception). 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 [Knill and Pouget 2004]. One striking observation from this work is the myriad ways in which human observers behave as near-optimal Bayesian observers. This observation, along with the behavioural and computational work on which it is based, has fundamental implications for neuroscience, particularly in how we conceive of neural computations and the nature of neural representations of perceptual variables [Knill and Pouget 2004].

Humans are clearly not optimal in the sense that they achieve the level of performance afforded by the uncertainty in the physical stimulus [Kozak and Castelo-Branco 2008]. Absolute efficiencies (a measure of performance relative to a Bayesian optimal observer) for performing high-level perceptual tasks are generally low and vary widely across tasks [Silva et al. 2008]. In some cases, this inefficiency is entirely due to uncertainty in the coding of sensory primitives that serve as inputs to perceptual computations; in others, it is due to a combination of sensory, perceptual and cognitive factors. The real test of the Bayesian coding hypothesis is in whether the neural computations that result in perceptual judgements or motor behaviour take into account the uncertainty in the information available at each stage of processing. Psychophysical work in several areas suggests that this is the case [Knill and Pouget 2004], being particularly evident in clinical models [Castelo-Branco, Mendes, Sebastião, Reis, Soares, Saraiva, Bernardes, Flores, Pérez-Jurado, and Silva 2007].

Several authors argue that these data strongly suggest that the brain codes complex patterns of sensory uncertainty in its internal representations and computations — see for example Knill and Pouget [2004], Pouget, Dayan, and Zemel [2000], Jacobs [2002], Rao [2004; 2005], Zemel, Dayan, and Pouget [1997], Denève, Latham, and Pouget [1999], Denève and Pouget [2004], Barber, Clark, and Anderson [2003].

On another perspective, the fact that there is strong evidence for a probabilistic computational framework in the human brain for perception, also brings forth the notion of *optimal percept*, or, in other words, that our percepts are our best guess as to what is in the world, given both sensory data and prior experience [Weiss, Simoncelli, and Adelson 2002, Geisler and Kersten 2002]. Such an "optimal guess" based on priors also suggests an explanation to why biological perception systems, when faced with perceptual scenarios which do not comply to the statistics of natural environments or when impaired due to disease or cerebral lesions, often fail to perceive the world as it is, substituting its exact description by the erroneous percepts called *perceptual illusions* — these are a direct result of perceptual ill-posed problems [Colas et al. 2010, Castelo-Branco 2005].

On the other hand, *interaction and navigation* requires maximal awareness of spatial surroundings, which in turn is readily obtained through active attentional and behavioural exploration of the environment. In animals with perception mainly based on visual sensing, visual, auditory and even tactile stimuli elicit gaze shifts (head and eye movements) to drive this active exploration.

For more than 20 years now, evidence has been accumulating from studies involving healthy human subjects that suggests parallel streams for visual processing for perception versus visual processing for the control of action [Dyde and Milner 2002]. In fact, in the human brain, mainly two pathways or streams, anatomically separate albeit interconnected in a complex fashion, have been found to be involved in sensory processing: the *dorsal pathway* and the *ventral pathway*.

Two main theories have arisen over the exact nature of the function of these two pathways, depending on whether emphasis is placed on the input distinctions or on output requirements. Over 20 years ago, Ungerleider and Mishkin [Ungerleider and Mishkin 1982, Mishkin, Ungerleider, and Macko 1983] described the functions of the two cortical systems based on the former, as a distinction between "object" versus "spatial" vision. Based on the latter, on the other hand, circa 10 years later, Goodale and Milner [1992] advanced the argument that the distinction is perhaps more parsimoniously described as one between visual "perception" and the visual control of "action". In this more recent account, the ventral stream of visual projections mediates the perception of objects and their relations, whereas the dorsal stream mediates the visual control of actions directed to these objects [Murphy, Carey, and Goodale 1998].

In either case, it is consensual that the dorsal stream, commonly called the "Where" or "How Pathway" depending on the theory, is associated with motion, representation of object locations, and control of the eyes and arms, especially when visual information is used to guide saccades or reaching, and that the ventral stream, commonly called the "What Pathway", is associated with form recognition and object representation. The latter is additionally believed to be associated with storage of long-term memory. It is also consensual nowadays that the widespread interconnections between the two pathways imply that their performances are strongly correlated in the undamaged brain (see, for example, Farivar [2009], Silver and Kastner [2009], Dyde and Milner [2002]), while decorrelation is more evident in clinical models of dorsal stream dysfunction (see Castelo-Branco et al. [2007]).

It is believed that there are multimodal perceptual feedback loops stemming from other sensory cortices and processing regions in the brain to the visual pathways. On the other hand, it is known that an additional phylogenetically older sensory processing site is heavily permeated by multimodal signals: the *superior colliculus* (SC).

In humans, the superior colliculus is involved in the generation of saccadic eye movements and eye-head coordination [Sparks 1999, Crawford, Ceylan, Klier, and Guitton 1999]. As with most larger vertebrates, sensory information that goes to the mesencephalon will be relayed via the thalamus to the cerebral cortex for interpretation, which, as has been mentioned previously, may be used to control the eyes in the dorsal stream. On the other hand, the SC can also mediate involuntary oculomotor movements without cortical involvement [Sparks 1999, Crawford et al. 1999]. However, when voluntary control is operating, then the *frontal eye fields* (FEF — the cortical analogue of the SC) mediate oculomotor behaviour.

The superior colliculus, also referred to as the *optic tectum* in other classes of vertebrates besides mammals, contains a map of visual space. The superficial layers of the SC receive a direct topographic projection from the contralateral retina. In addition, both superficial and deep layers of the SC receive indirect retinotopic maps from descending cortical projections and possibly from intrinsic SC connections. As a result of these connections, SC units are excited by visual stimuli in a restricted region of the visual field (receptive field) and are inhibited by stimuli located outside of this region. The receptive fields are organized systematically across the surface of the SC to form a visual map that represents contralateral space only and stops at the representation of the vertical meridian (as opposed to the optic tectum, which represents the entire visual field of the contralateral retina) [Knudsen and Brainard 1995].

In the deep layers of the SC (and also in the superficial layers of the tectum in the barn owl), units respond to auditory stimuli as well. Auditory units also exhibit excitatory receptive fields that are surrounded by inhibitory regions, just like visual neurons. Auditory receptive fields are larger than visual receptive fields. In the barn owl, a nocturnal predator that relies heavily on sound localisation to capture prey, auditory receptive fields average approximately 30° in azimuth and 50° in elevation compared with visual receptive fields, which average only about 12° in diameter. In other species, such as cats and monkeys, auditory receptive fields are generally even larger and for many neurons responses can be elicited by sounds located within an entire hemifield [Knudsen and Brainard 1995].

However, even for neurons with large receptive fields, the strength of response usually varies with the location of auditory stimuli so that strong responses are only elicited when sounds arise from a much more restricted region of space, the *auditory best area*. Just as with visual receptive fields, auditory best areas are systematically organised across the SC, thereby creating a map of auditory space. The map of auditory space (i.e. tonotopic representation), which is based on the tuning of neurons to sound localisation cues, is approximately aligned with the map of visual space in the SC. A close correspondence exists between the centres of auditory best areas and visual receptive fields of neurons encountered in penetrations perpendicular to the SC surface. Moreover, many neurons in the intermediate and deep layers of the SC receive convergent auditory and visual information. These neurons have closely aligned auditory and visual receptive fields and thus form a bimodal space map [Knudsen and Brainard 1995].

In summary, the SC receives visual, as well as auditory, inputs in its superficial layers, and the deeper layers of the colliculus are connected to many sensorimotor areas of the brain [Knudsen and Brainard 1995, Wallace, Meredith, and Stein 1998]. The colliculus as a whole is thought to help orient the head and eyes towards salient stimuli [Sparks 1999, Crawford et al. 1999]. It contains a retinotopic visuoauditory spatial map with polar configuration.

The spatial location of an object may be, in principle, represented with reference to two fundamental classes of spatial coordinate frames: *egocentric* and *allocentric*. In the egocentric frames, the position of objects is encoded with reference to the body of the observer or, more specifically, to relevant body parts, such as the head, trunk, and/or arm. Egocentric representations of objects may be used for the organization of goal-directed movements, such as reaching a target or avoiding a dangerous stimulus. In the allocentric coordinate frames, by contrast, objects are primarily represented with reference to their spatial and configurational properties, such as the relationships among their different component parts and among different objects in the environment. Representations encoding the configurational properties of objects may be useful for their identification. Objects, in ecological conditions, are typically seen from a variety of egocentric (observer-based) perspectives, suggesting a close interaction between body- and object-based reference frames [Galati, Lobel, Vallar, Berthoz, Pizzamiglio, and Bihan 2000].

In contrast to the perception of spatial layout, provided by the ventral stream, the computation of spatial location, carried out by the dorsal stream, is entirely related to the guidance of specific visuomotor actions, such as grasping an object, locomoting around obstacles, or gazing at different objects in a scene. As a consequence, the dorsal stream mechanisms, as with the superior colliculus, do not compute the allocentric location of a target object, i.e. its location relative to other objects in the scene, but rather the *egocentric* coordinates of the location of the object with respect to the observer [Murphy et al. 1998].

In short, both dorsal and ventral visual systems compute information about spatial location, but in very different ways: allocentric spatial information about the layout of objects in the visual scene is computed by the ventral stream mechanisms, which mediate perception, while precise egocentric spatial information about the location of an object in a body-centred frame of reference is computed by the dorsal stream mechanisms, which mediate the visual control of action [Murphy et al. 1998]. On the other hand, direction and distance in egocentric representations are believed to be separately specified by the brain [Gordon, Ghilardi, and Ghez 1994, McIntyre, Stratta, and Lacquaniti 1998].

The question of error in visual perception of egocentric distance (i.e. depth away from the observer) has been a hot topic of discussion for several years now [Cutting and Vishton 1995]. An important fact about results from distance judging experiments is that mean egocentric depth (distance away from the observer) is systematically foreshortened when compared to frontal depth (distances extended laterally in front of the observer, orthogonal to a given line of sight); indeed, many would suggest that such judgements would be foreshortened still further. Direct scaling and related methods are often criticized as being open to "cognitive correction;" most adults know that distances foreshorten as they increase and could easily compensate judgements with this knowledge [Cutting and Vishton 1995].

In any case, it is reasonably consensual that the shortcomings of absolute or quasi-absolute visual depth cues can be compensated, either by top-down influences, as stated above, or even by other visual depth cues, in both 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) and “action space” (circular region just beyond personal space, which may be acted upon reasonably quickly, directly or indirectly, by the observer — the utility of disparity and motion perspective decline to an effective threshold value of 10% at about 30 m, and thus Cutting and Vishton [1995] claim this effectively provides the outer boundary of this space), or perhaps even a bit farther, while the foreshortenings become virtually unavoidable beyond this point (the so-called “vista space” [Cutting and Vishton 1995]). Distance or depth errors are apt to occur in distance portions of the visual field because cues of depth are attentuated or are below threshold and therefore are unable to support the perception of depth between distant objects at different positions [Cutting and Vishton 1995]. So, even more important than the actual perceptual underestimation of depth become just-discriminable depth thresholds, which 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 [Cutting and Vishton 1995].

These findings support the construction of a framework that allows fast processing of perceptual 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 (as in the ventral stream) to other stages of processing — this would be analogous to a tennis player being required to hit a ball regardless of perception of its texture properties. This framework might be considered as bearing the spherical (i.e. coding 3D distance and direction) spatial configuration counterpart of the dorsal stream/superior colliculus egocentric representations in the brain. Moreover, the idea of constructing a short-term perceptual memory performing efficient, lossless compression through log-partitioning of depth seems to be reasonably supported by human depth perception and the just-discriminable depth thresholds phenomenon.

In conclusion, in this text we propose a bioinspired perceptual model with focus on Bayesian visuoauditory integration supported by proprioception (including vestibular sensing) that serves as a short-term spatial memory framework for active perception and also sensory control of action, with no immediate interest in object perception. The computational models described herewith will support the construction of a simultaneously flexible and powerful robotic implementation to be used in real-world applications, such as human-machine interaction or mobile robot navigation.

1.2 Related Work

Our goals, stated in the last sentence of the previous text, imply: the use of an accurate representation of metric space for multisensory-based perception, with support for the analysis of the temporal evolution of the occupation of that space; the use of that representation with the purpose of linking perception to action (both in the sense on acting upon the environment itself, but also on acting to redirect the sensors so as to make perception a dynamic process).

Fusing computer vision, binaural sensing and vestibular sensing using a unified framework, to the author's knowledge, has never been addressed. In fact, the extension of the well-known probabilistic occupancy grid model to an egocentric, log-spherical configuration as a solution to problems remotely similar to the ones presented in this text is also unprecedented, as far as is known by the author.

Metric maps are very intuitive, yield a rigorous model of the environment and help to register measurements taken from different locations. Grid-based maps — the most popular metric maps in mobile robotics applications — have been widely used to represent the environments' geometry through sets of polyhedra. Although metric maps allow for the storage of a high level of detail of the environment, their main shortcoming is that they are not particularly not well suited for exploration, since they do not scale gracefully as the surveying region increases in size [Rocha 2005]. They are, on the other hand, of extreme usefulness to promote direct action, such as in manipulation or obstacle avoidance tasks, where precise size and time-to-collision estimates are needed.

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 independent between each other [Moravec and Elfes 1985, Moravec 1988, Elfes 1989, Pagac, Nebot, and Durrant-Whyte 1998]. This geometric model has been extensively used in robotics mainly due to its simplicity and suitability for decision-theoretic approaches. 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.

The main hypothesis of occupancy 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.

This assumption, however advantageous it may be to achieve close-to-real-time performances, is not without its drawbacks: more specifically, there may be a trade-off in precision, specially when considering sensor readings that affect several adjacent cells. Nevertheless, this trade-off has been found in practice to be mostly irrelevant given the requirements of the majority of applications.

Rocha, Dias, and Carvalho [2005a], Rocha [2005] introduced an upgraded version of the occupancy map first introduced by Stachniss and Burgard [2003], wherein the notion of occupancy grid was augmented in order to avoid a strictly binary representation of each cell's occupancy (i.e., free or occupied) through the replacement of these representations with continuous random variables encoding the *percentage of occupancy* of that cell (dubbed "cell coverage" by the authors), refining it even further through the use of probability distributions over the occupancy percentage values. They integrated a simple implementation of Bayesian filtering to address scene dynamics and implement temporal consistency.

More recently, Coué, Pradalier, Laugier, Fraichard, and Bessière [2006] and Tay, Mekhnacha, Chen, Yguel, and Laugier [2007] expanded on the occupancy grid by explicitly introducing *Bayesian filtering*. These versions of the occupancy grid inherit the advantages of the original, adding to it a further advantage, by modelling the dynamics of the environment and by enforcing robustness relative to object occlusions through the use a novel two-step mechanism which permits taking the sensor observations history and the temporal consistency of the scene into account.

This approach is derived from the Bayesian filtering approach, which explains why these operations together with the occupancy maps are called by the authors *Bayesian Occupancy Filters* (BOF). The differences between the model by Coué et al. and by Tay et al. rely on the fact that the former use a compact 4D formulation to store the information regarding 2D position and 2D velocity estimates which allows the representation of overlapping objects, while the latter use a 2D formulation which allows for the inference of velocity distributions.

Bayes filters [Jazwinsky 1970] address the general problem of estimating the state sequence $x^k, k \in \mathbb{N}$ of a system given by:

$$x^k = f^k(x^{k-1}, u^{k-1}, w^k) \tag{1.1}$$

where f^k is a possibly nonlinear transition function, u^{k-1} is a "control" variable (e.g. speed or acceleration) for the sensor which allows to estimate its egomotion between time $k-1$ and time k, and w^k is the process noise. This equation describes a Markov process of order one.

Let z^k be the sensor observation of the system at time k. The objective of the

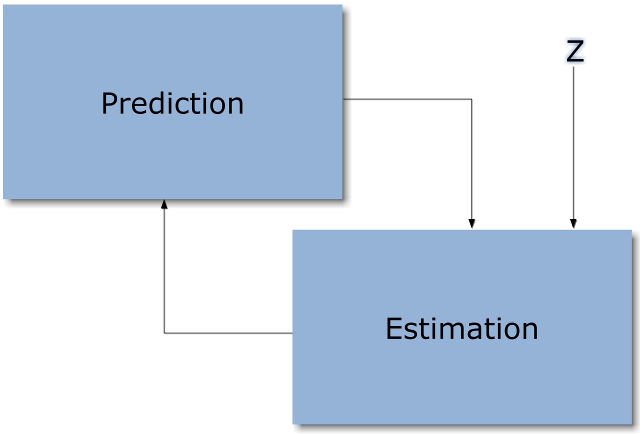


Figure 1.2: The Bayesian Occupancy Filter as a recursive loop.

filtering is to recursively estimate x^k from the sensor measurements:

$$z^k = h^k(x^k, v^k) \tag{1.2}$$

where h^k is a possibly nonlinear function and v^k is the measurement noise. This function models the uncertainty of the measurement z^k of the system's state x^k.

In other words, the goal of the filtering is to recursively estimate the probability distribution $P(X^k|Z^k)$, known as the *posterior distribution*. In general, this estimation is done in two stages: *prediction* and *estimation*. The goal of the prediction stage is to compute an *a priori* estimate of the target's state known as the *prior distribution*. The goal of the estimation stage is to compute the *posterior distribution*, using this *a priori* estimate and the current measurement reading of the sensor.

Exact solutions to this recursive propagation of the posterior density do exist in a restrictive set of cases. In particular, the Kalman filter [Kalman 1960, as cited by Welch and Bishop 2006] is an optimal solution when functions f^k and h^k are linear and the noise terms w^k and v^k are Gaussian. But, in general, solutions cannot be determined analytically and an approximate solution has to be computed.

In this case, the state of the system is given by the occupancy state of each cell of the grid, and the required conditions for being able to apply an exact solution such as the Kalman filter are not always verified. Moreover, the particular structure of the model (occupancy grid) and the real-time constraint coming from most robotic applications, leads to the development of the concept of the Bayesian Occupancy Filter. This filter consists of estimating the occupancy state using the Bayesian filter two-step mechanism, as depicted in Fig. 1.2.

In our specific application domain, where a 3D metric and egocentric representation is required, common occupancy 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 such a configuration known to the author was presented by Zapata, Jouvencel, and Lépinay [1990] — the advantages of using such a representation directly in robot navigation are clearly stated therewith, in particular in motion planning for fast mobile robots.

2. Implementation-wise, regular partitioning in Euclidean space, while still manageable in 2D, renders temporal performances impratical 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, as mentioned above, 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 [2005].

Active perception has been an object of study in robotics for decades now, specially active vision, which was first introduced by Bajcsy [1985] and later explored by Aloimonos, Weiss, and Bandyopadhyay [1987]. Many perceptual tasks tend to be simpler if the observer is active and controls its sensors [Aloimonos et al. 1987]. Active perception is thus an intelligent data acquisition process driven by the measured, partially interpreted scene parameters and their errors from the scene. The active approach has the important advantage of making most ill-posed perception tasks tractable [Aloimonos et al. 1987].

Recent work in active vision by Tsotsos and Shubina [2007] and Bohg, Barckholst, Huebner, Ralph, Rasolzadeh, Song, and Kragic [2009], the former for target search and the latter for object grasping, contrary to our solution use an explicit representation for objects to implement active perception. On the other hand, several solutions for target applications similar to ours avoid explicit object representation by resorting to a bottom-up saliency approach such as defined by Itti, Koch, and Niebur [1998] — examples of these would be Shibata, Vijayakumar, Conradt, and Schaal

[2001], Breazeal, Edsinger, Fitzpatrick, and Scassellati [2001] and Dankers, Barnes, and Zelinsky [2007]. Finally, Dankers, Barnes, and Zelinsky [2005] use an approach similar to ours, with an egocentric three-dimensional occupancy grid for integrating range information using active stereo and a Bayesian approach, also detecting 3D mass flow. However, this solution suffers from the downside of using an Euclidean tesselation of space, which complicates sensor models for map updating and fixation computation due to the compulsory use of ray-tracing methods. These works, as most solutions in active perception, use a behavioural approach; an alternative is a probabilistic approach that attempts to reduce uncertainty on a part of the world state, modelled as belief [de Croon, Sprinkhuizen-Kuyper, and Postma 2009]. Our work intends to combine both variants into a coherent, albeit more powerful approach.

Active multisensory perception using spatial maps has, contrastingly, been the object of study since only much recently. Few other explicit models exist, although many artificial perception systems include some kind of simple attention module that drives gaze towards salient auditory features. As an example of a full-fledged multisensory attention model, Koene, Morén, Trifa, and Cheng [2007] present a general architecture for the perceptual system of a humanoid robot featuring multisensory (audiovisual) integration, bottom-up salience detection, top-down attentional feature gating and reflexive gaze shifting, which is of particular relevance to our work. The complete system focuses on the multisensory integration and desired gaze shift computation performed in the "Superior Colliculus (SC)" module [Koene et al. 2007]. This allows the robot to orient its head and eyes so that it can focus its attention on audio and/or visual stimuli. The system includes mechanisms for bottom-up stimulus salience based gaze/attention shifts (where salience is a function of feature contrast) as well as top-down guided search for stimuli that match certain object properties. In order to facilitate interaction with dynamic environments the complete perceptual-motor system functions in real-time [Koene et al. 2007].

As with Koene et al. [2007], our solution implements active visuoauditory perception, adding to it vestibular sensing/proprioception so as to allow for sensor fusion given a rotational egomotion. However our solution differs from purely saliency-based approaches in that it also implements an active exploration behaviour based on the entropy of the occupancy grid, so as to promote gaze shifts to regions of high uncertainty.

1.3 Contributions of This Work and Structure of the Dissertation

Our work will contribute in providing a rather complete framework for active multimodal perception — introducing an approach which, while not strictly neuromimetic, finds its roots in the role of the dorsal perceptual pathway of the human brain — that, as mentioned above, will support the construction of a simultaneously flexible and powerful robotic implementation to be used in real-world applications, such as human-machine interaction or mobile robot navigation.

Its main strength lies on the fact that it offers a solution which is naturally fitting for acting upon the environment and also for the integration of readings from multiple sensors, since both processes inherently depend on egocentric reference frames.

After presenting our general motivations in this chapter, we mainly expect to demonstrate our contributions in the remainder of this dissertation (which were also reported in the publications listed in Appendix C and referred to in the items below), comprising:

- The development of a Bayesian framework to support active multimodal perception research (chapter 2 and Ferreira, Bessière, Mekhnacha, Lobo, Dias, and Laugier [2008a], Ferreira, Pinho, and Dias [2008c], Pinho, Ferreira, Bessière, and Dias [2008], Ferreira, Pinho, and Dias [2008b]) which:
    - Allowed the definition of a coherent experimental paradigm.
    - Promoted a batch of preliminary baseline studies of human visuoauditory motion perception.
    - Spurred the construction of a robotic experimental platform for active multimodal perception.
    - Deals with perceptual uncertainty and ambiguity using a novel spatial configuration for an occupancy grid representation, offering some adaptive ingredients that would form a reasonable bioinspired basis for a full-fledged robotic perception system.
    - Offers efficient, robust and novel probabilistic solutions for cyclopean geometry-based stereovision and auditory perception based only on binaural cues.
    - Deals with sensor fusion in a natural way, consistent with most of the inherent properties of sensation.

  - Allows for fast processing of perceptual inputs to build a spatial representation that promotes immediate action on the environment, both for active perception, and for manipulation and navigation purposes.

- The design and execution of baseline experiments of human multimodal motion perception (chapter 3), namely:
  - A baseline experiment to determine the influence of horizontal-vertical anisotropies on visual motion perception.
  - A baseline experiment to determine the influence of visual and auditory context when the auditory and visual modalities are attended to, respectively.

- The development of a GPU-based implementation of a real-time solution for active exploration using the aforementioned framework, capitalising on the potential for parallel computing of most of its algorithms (chapter 4 and Ferreira et al. [2008b; 2009a], Lobo et al. [2009], Ferreira et al. [2009b; 2010]).

- The implementation of a hierarchical Bayesian active perception system that simulates several bottom-up-driven human behaviours (chapter 5 and Ferreira and Dias [2010]), exhibiting the following desirable properties:

  Emergence — High-level behaviour results from low-level interaction of simpler building blocks.

  Scalability — Seamless integration of additional inputs is allowed by the Bayesian Programming formalism used to state the models of the framework.

  Adaptivity — Initial "genetic imprint" of distribution parameters may be changed "on the fly" through parameter manipulation, thus allowing for the implementation of goal-dependent behaviours (i.e. top-down influences).

A supporting website — `http://paloma.isr.uc.pt/~jfilipe/BayesianMultimodalPerception` — was developed as a companion to this text, for dissemination purposes. Due to the dynamical nature of the phenomena studied herewith and of the perceptual framework developed throughout this work, referral to the animations and videos available online at this site is recommended.

Chapter 2

A Bayesian Framework for Active Multimodal Perception Research

2.1 An Experimental Paradigm for Multimodal Perception Research

As depicted on Fig. 2.1 on the following page, the objectives of our work are, in general terms, to experimentally study biological (human) models of integrated visual, auditory and vestibular perception, apply them to artificial platforms with artificial sensors performing the same tasks, test them, and ultimately revise the initial models in a feedback fashion, all this under a Bayesian modelling framework, ultimately following a process inspired by the *Bayesian ideal observer analysis* as defined by Geisler [1989a;b; 2003] and applied to perceptual model development by Schrater and Kersten [2001].

To support the attainment of this objective, we believe that the availability of a unified framework for experimental procedures is essential. A way of achieving this consistency is to carefully delineate the interactions between the experimental techniques, the models, the inputs to the system and the outputs provided by the techniques, ultimately relating them within a timeline, as shown on Figure 2.2.

Given its *input* → *system* → *output* nature, this framework can be easily extended to experiments involving artificial observers (in fact, the ideal observer is already an artificial observer), thus providing a unified solution for the experiments to be conducted.

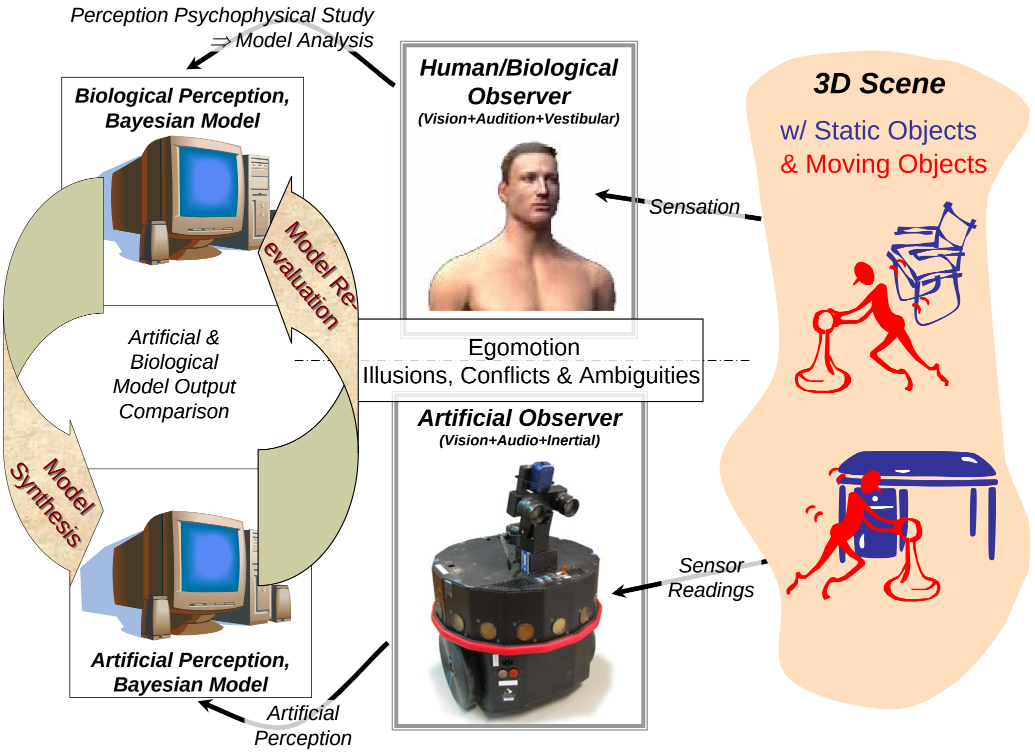

Perception Psychophysical Study
⇒ Model Analysis
Human/Biological Observer
(Vision+Audition+Vestibular)
3D Scene
w/ Static Objects
& Moving Objects
Biological Perception, Bayesian Model
Model Re-evaluation
Sensation
Artificial & Biological Model Output Comparison
Egomotion
Illusions, Conflicts & Ambiguities
Artificial Observer
(Vision+Audio+Inertial)
Model Synthesis
Sensor Readings
Artificial Perception, Bayesian Model
Artificial Perception

Figure 2.1: Experimental paradigm for multimodal perception research.

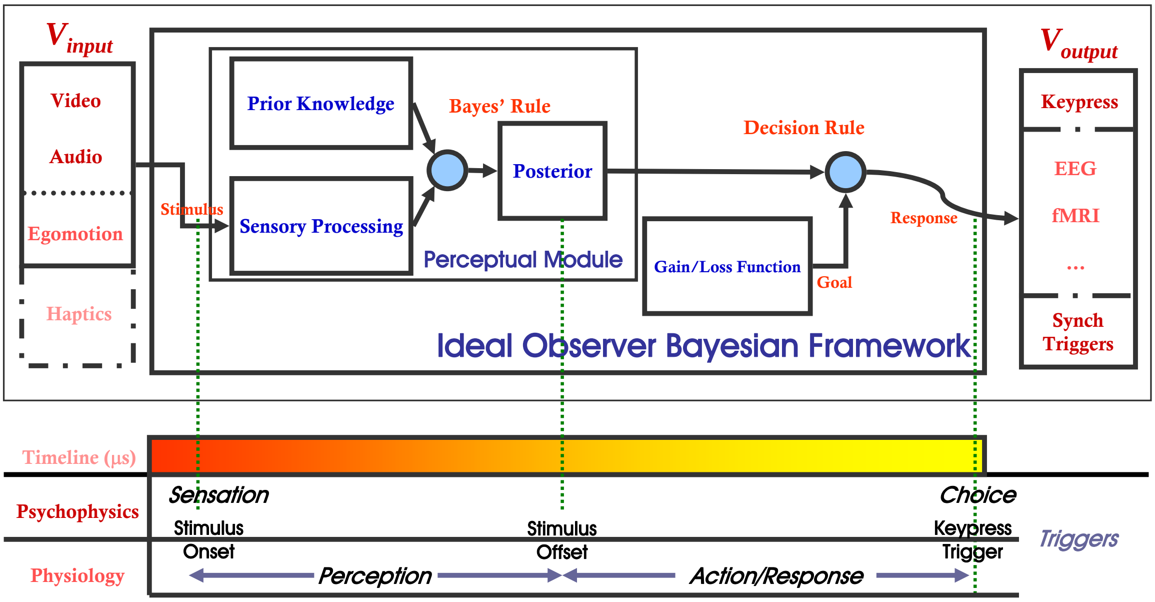


Figure 2.2: A unified framework for experimental procedures. The ideal observer (i.e., the tentative model) is interchangeable with a human observer (i.e., the subject under study) and performances can thus be compared. As can be seen in the schematic, experimental techniques may include, but are not limited to, psychophysics, electrophysiology, fMRI, body tracking, etc.

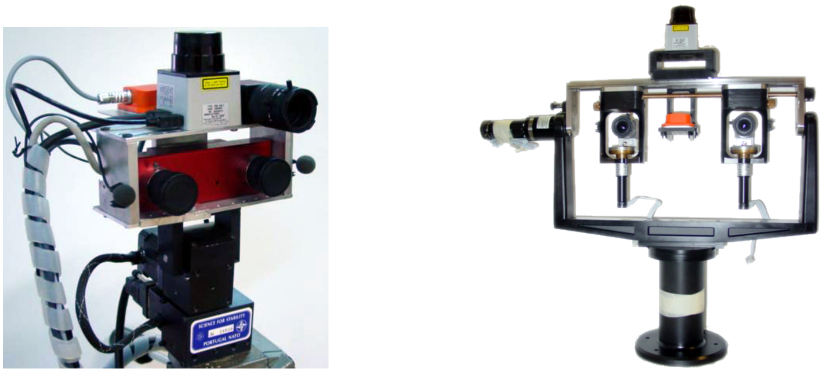


Figure 2.3: View of the first version of the Integrated Multimodal Perception Experimental Platform (IMPEP), on the left, and current version on the right. The latter added vergence capabilities to the stereovision system besides improved motor control and conditioning. The active perception head mounting 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).

2.2 The Integrated Multimodal Perception Experimental Platform

2.2.1 Platform description

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. 2.3.

In the current version of the platform, 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. Initial pitch and roll position are taken from the initial moment with the sensor at rest using the gravity acceleration.

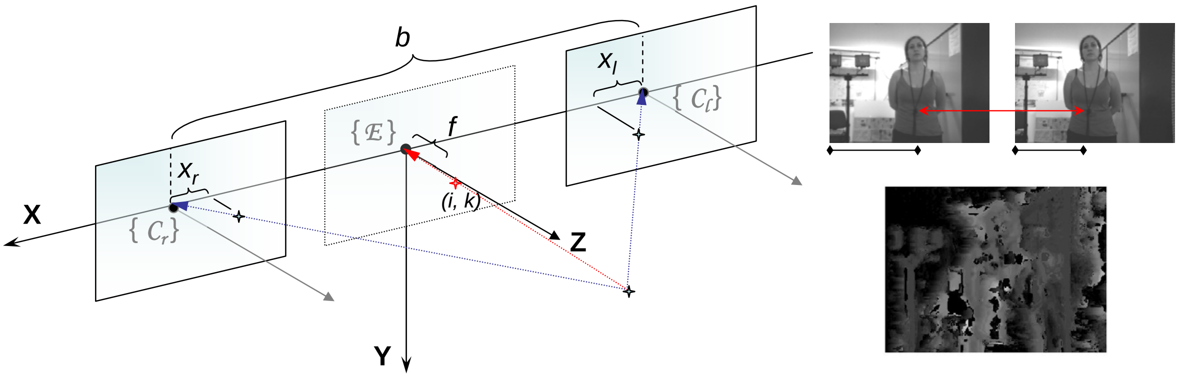


Figure 2.4: Cyclopean geometry for stereovision. The use of cyclopean geometry (pictured on the left for an assumed frontoparallel configuration) allows direct use of the egocentric reference frame for depth maps taken from the disparity maps yielded by the stereovision system (of which an example is shown on the right).

2.2.2 Sensory processing

Vision system

As mentioned in the introductory chapter, several authors argue that current evidence strongly suggests that the brain codes complex patterns of sensory uncertainty in its internal representations and computations. One such representation is believed to be neural population coding (e.g., average firing rate) — see for example Knill and Pouget [2004], Pouget et al. [2000], Jacobs [2002], Rao [2005], Zemel et al. [1997], Denève et al. [1999], Barber et al. [2003].

Our motivations suggest a tentative data structure analogous to neuronal population activity patterns to represent uncertainty in the form of probability distributions [Pouget et al. 2000]. Thus, a spatially organised 2D grid may have each cell (corresponding to a virtual photoreceptor in the cyclopean view — see Fig. 2.4) associated to a “population code” extending to additional dimensions, yielding a set of probability values encoding a N-dimensional probability distribution function or pdf (see Fig. 2.5).

The stereovision algorithm used with the first version of the IMPEP head was an adaptation of the fast and simple coherence detection approach by Henkel [1998; 2000], which is easily converted from its deterministic nature into a probabilistic implementation simulating the population code-type data structure. The workings of the algorithm and our own adaptation are described in the following lines.

Disparity estimators in real biological networks can and will vary in various properties, notably in the separation of the centre of their receptive fields in the left and

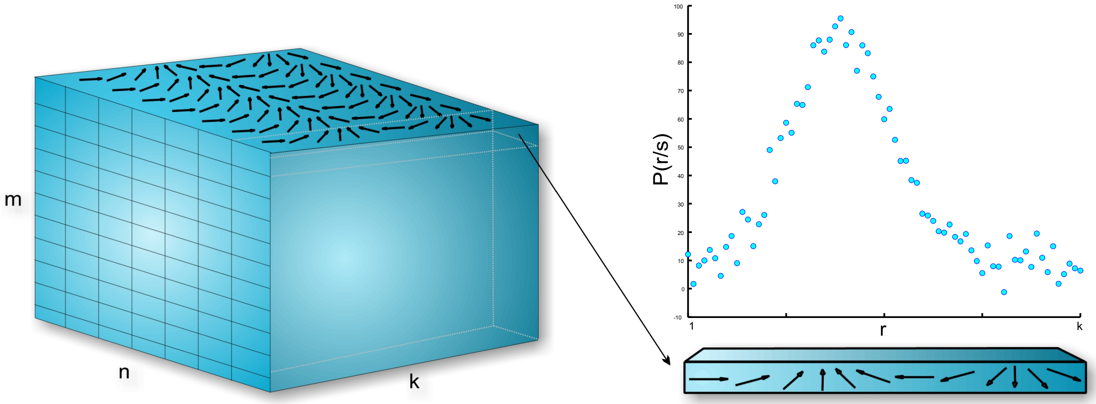


Figure 2.5: Population code data structure. On the left, a spatially organised 2D grid has each cell (which might correspond, for example, to a specific area on the retina or a pixel on a digital image) associated to a population code simulation extending to a third dimension, represented on the right — i.e. a set of probability values of a neuronal population encoding a pdf (in this example, for preferred directions). Note that this map does not precisely mimic the cortical columnar architecture, and is just an approximation, and that the pdf can in fact extend to more than a single dimension (e.g., if the encoded property would be local velocity, two dimensions would be necessary so as to represent speed and direction).

right eye. Other possible parameters of interest are the spatial orientation, the scale or the phase of the Gabor filter patches.

In the simplified network by Henkel, any type of disparity estimator can be used in different spatial scales[1]. At a single scale, image data is fed diagonally into layers of identical disparity units. This creates layers of units having a common and fixed separation of receptive fields in the left and right eye. Disparity units stacked vertically above each other sample space in a common view direction, and it is here where the coherence detection scheme sets in. Disparity units operating at different scales are simply included in the appropriate disparity stacks — all disparity units i in a stack will have different, but slightly overlapping, working ranges $D_i = [\delta_i^{min}, \delta_i^{max}]$ for valid disparity estimates. An object with true disparity δ, seen in the common view direction of the stack, will therefore split the stack into two disjunct classes: the class $\mathcal{C}$ of estimators with $\delta \in D_i$ for all $i \in \mathcal{C}$, and the rest of the stack, $\bar{\mathcal{C}}$, with $\delta \notin D_i$. All disparity estimators. All disparity estimators $\in \mathcal{C}$ will code more or less the true

[1] The stereovision setup is assumed to be in frontoparallel configuration, or the stereo images rectified to simulate such a configuration.

disparity $\hat{\delta}_i \approx \delta$, but the estimates of units belonging to $\bar{\mathcal{C}}$ will be subject to random aliasing effects, depending in a complicated way on image content and disparity range D_i of the unit.

We will thus have $\hat{\delta}_i \approx \delta \approx \hat{\delta}_j$ whenever units i and j belong to $\mathcal{C}$, and random relationships otherwise. A simple coherence detection within each stack, i.e., searching for the largest cluster of units within $|\hat{\delta}_i - \hat{\delta}_j| < \epsilon$ similarity measure[2], will be sufficient to single out $\mathcal{C}$. The true disparity δ in the view direction of the stack can be simply estimated as an average over all coherent coding units

$$\hat{\delta} = \left\langle \hat{\delta}_i \right\rangle_{i \in \mathcal{C}} \tag{2.1}$$

which can be then used to construct a disparity map that can then be used to estimate depth (i.e., 3D structure) by using the camera's extrinsic parameters estimated during calibration. Coherence maps are also constructed by calculating a simple verification count derived from the relative number of coherently acting disparity units in each stack n, i.e. by calculating the ratio

$$\lambda(k,i) = c_n = \frac{|\mathcal{C}|}{|\mathcal{C} \cup \bar{\mathcal{C}}|} \tag{2.2}$$

where $|\bullet|$ denotes set cardinality (i.e. it is, in fact, an element count operator); collecting all c_n and ordering them by projection line, a coherence map composed of confidence values $\lambda(k,i)$ for each pixel (k,i) of the left image can be constructed. Probability distribution functions corresponding to each projection line are then assembled using values within this coherence map (their inverses are effectively uncertainty measures that can be used to derive distribution standard deviations/variances) together with the disparity estimates (used as expected values in the distributions) so as to build the population code-type data structure with pdfs defined as likelihood functions; no further assumption on the actual type of distributions is made at this stage.

The immediate advantage of such an implementation would be the availability of confidence measures per depth measurement taken from the variance of the corresponding pdf. A further advantage would be the possibility of future use of powerful probabilistic/belief propagation methods to allow for temporal integration of several frames. Finally, the probabilistic nature of such an algorithm would allow effortless seeming with higher-level Bayesian cue integration modules.

To conclude, the cyclopean view of the stereovision system can also be easily derived using this algorithm. Let I_i^L and I_i^R be the left and right input data of disparity unit

[2] The actual threshold value ϵ is chosen by the authors empirically as being between the 0.3 and 0.5, although they claim that the exact value is not critical for performance.

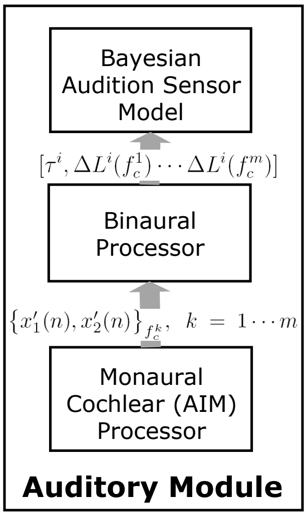

Bayesian Audition Sensor Model
$[\tau^i, \Delta L^i(f_c^1) \cdots \Delta L^i(f_c^m)]$
Binaural Processor
$\{x'_1(n), x'_2(n)\}_{f_c^k}, \; k = 1 \cdots m$
Monaural Cochlear (AIM) Processor
Auditory Module

Figure 2.6: The IMPEP Bayesian binaural system.

i, respectively. A simple average over the coherently coding disparity units

$$I^{\mathcal{C}} = \left\langle I_i^L + I_i^R \right\rangle_{i \in \mathcal{C}} \tag{2.3}$$

gives the image intensities of the cyclopean view.

Auditory system

The Bayesian binaural system presented herewith is composed of three distinct and consecutive processors (Fig. 2.6): 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.

The first stages of auditory processing consist of cochlear and auditory periphery processing, which produces what is called an *auditory image model* (AIM) [Patterson et al. 1995]. The AIM processor implements a functional model of a cochlea that simulates the phase-locked activity that complex sounds produce in the auditory nerve.

Spectral analysis, the first stage of the AIM, is performed by a bank of auditory filters which converts each digitised wave that composes the stereo signal into an array of filtered waves. This processing is done using *gammatone filters* [de Boer 1975, Immerseel and Peeters 2003], linearly distributed along a frequency scale measured in *equivalent rectangular bandwidths* (ERBs), as defined by [Patterson, Holdsworth, Nimmo-Smith, and Rice 1988] for simulating the cochlea, obtaining a model of *basilar membrane motion* (BMM) through frequency band decomposition.

The second stage of the AIM simulates the mechanical/neural transduction process performed by the inner haircells. It converts the BMM into a *neural activity pattern* (NAP), which is the AIM's representation of the afferent activity in the auditory nerve [Patterson, Allerhand, and Giguère 1995]. In this stage the envelopes of the signals are first compressed, and then subjected to halfwave rectification followed by a squaring and lowpass filtering, resulting in m stereo audio signal pairs corresponding to m frequency channels with respective frequency centre f_c^k, $\left\{x_1'(n), x_2'(n)\right\}_{f_c^k}$, $k = 1 \cdots m$.

Sound waves arising from a source on our left will arrive at the left ear first. This small, but perceptible, difference in arrival time (known as an ITD, interaural time difference) is an important localisation cue and is detected by the *inferior colliculus* in primates, which acts as a temporal correlation detector array, after the auditory signals have been processed by the cochlea. Similarly, for intensity, the far ear lies in the head's "sound shadow", giving rise to interaural level differences (ILDs) [King, Schnupp, and Doubell 2001, Kapralos, Jenkin, and Milios 2003]. ITDs vary systematically with the angle of incidence of the sound wave relative to the interaural axis, and are virtually independent of frequency, representing the most important localisation cue for low frequency signals ($< 1500\,\mathrm{Hz}$ in humans). ILDs are more complex than ITDs in that they vary much more with sound frequency. Low-frequency sounds travel easily around the head, producing negligible ILDs. ILD values produced at higher frequencies are larger, and are increasingly influenced by the filter properties of each external ear, which imposes peaks and notches on the sound spectrum reaching the eardrum.

Moreover, when considering sound sources within $1 - 2$ meters of the listener, binaural cues alone can even be used to fully localise the source in 3D space (i.e. azimuth, elevation and distance). Iso-ITD surfaces form hollow cones of confusion with a specific thickness extending from each ear in a symmetrical configuration relatively to the medial plane. On the contrary, iso-ILD surfaces, which are spherical surfaces, delimit hollow spherical volumes, symmetrically placed about the medial plane and centred on a point on the interaural axis [Shinn-Cunningham, Santarelli, and Kopco 2000]. Thus, for sources within 2 meters range, the intersection of the ILD and ITD volumes is a torus-shaped volume [Shinn-Cunningham et al. 2000]. If the source is more than 2 meters away, the change in ILD with source position is too gradual to provide spatial information (at least for an acoustically transparent head), and the source can only be localised inside a volume within the cone of confusion delimited by the respective iso-ITD surfaces [Shinn-Cunningham et al. 2000].

Given this background, we have decided to adapt the solution by Faller and Merimaa [Faller and Merimaa 2004a] to implement the binaural processor. Using this algorithm, interaural time difference and interaural level difference cues are only considered at time instants when only the direct sound of a specific source has nonnegligible energy in the critical band and, thus, when the evoked ITD and ILD represent the direction of that source (corresponding to the process involving the *superior olivary complex* (SOC) and the *central nucleus of the inferior colliculus* (ICc) in mammals). They show how to identify such time instants as a function of the *interaural coherence* (IC). The source localisation suggested by the selected ITD and ILD cues are shown to imply the results of a number of published psychophysical studies related to source localisation in the presence of distractors, as well as in precedence effect conditions [Zurek 1987]. This algorithm thus amplifies the signal-to-noise ratio and facilitates auditory scene analysis for multiple auditory object tracking, and is briefly summarised in the following paragraphs — for more details, please refer to [Faller and Merimaa 2004a].

The ITD and IC, denoted respectively by $\tau(n)$ and $c_{12}(n)$, where n indexes the sample currently being processed, are estimated from the normalised cross-correlation functions of the signals from left and right ear for each centre frequency f_c, respectively x'_1 and x'_2. The normalisation of the cross-correlation function is introduced in order to get an estimate of the IC, defined as the maximum value of the instantaneous normalised cross-correlation function. This estimate describes the coherence of the left and right ear input signals. In principle, it has a range of $[0; 1]$, where 1 occurs for perfectly coherent x'_1 and x'_2. However, due to the DC offset of the halfwave rectified signals, the values of c_{12} are typically higher than 0 even for independent (nonzero) x'_1 and x'_2. Thus, the effective range of the interaural coherence c_{12} is compressed to $[a; 1]$ by the neural transduction. The compression is more pronounced (larger a) at high frequencies, where the low pass filtering of the half-wave rectified critical band signals yields signal envelopes with a higher DC offset than in the signal wave forms [Faller and Merimaa 2004a].

The ILD, denoted as $\Delta L(n)$, is then computed using the signal levels at the corresponding offsets [Faller and Merimaa 2004a]. Note that due to the envelope compression the resulting ILD estimates will be smaller than the level differences between the ear input signals. For coherent ear input signals with a constant level difference, the estimated ILD (in dB) will be 0.23 times that of the physical signals [Faller and Merimaa

2004a].

When several independent sources are concurrently active in free field, the resulting cue triplets $\{\Delta L(n), \tau(n), c_{12}(n)\}$ can be classified into two groups [Faller and Merimaa 2004a]: (1) Cues arising at time instants when only one of the sources has power in that critical band. These cues are similar to the free-field cues — localisation is represented in $\{\Delta L(n), \tau(n)\}$, and $c_{12}(n) \approx 1$. (2) Cues arising when multiple sources have non-negligible power in a critical band. In such a case, the pair $\{\Delta L(n), \tau(n)\}$ does not represent the direction of any single source, unless the superposition of the source signals at the ears of the listener incidentally produces similar cues. Furthermore, when the two sources are assumed to be independent, the cues are fluctuating and $c_{12}(n) < 1$. These considerations motivate the following method for selecting ITD and ILD cues. Given the set of all cue pairs, $\{\Delta L(n), \tau(n)\}$, only the subset of pairs is considered which occurs simultaneously with an IC larger than a certain threshold, $c_{12}(n) > c_0$. This subset is denoted

$$\{\Delta L(n), \tau(n) | c_{12}(n) > c_0\} \tag{2.4}$$

The same cue selection method is applicable for deriving the direction of a source while suppressing the directions of one or more reflections. When the "first wave front" arrives at the ears of a listener, the evoked ITD and ILD cues are similar to the free-field cues of the source, and $c_{12}(n) \approx 1$. As soon as the first reflection from a different direction arrives, the superposition of the source signal and the reflection results in cues that do not resemble the free-field cues of either the source or the reflection. At the same time IC reduces to $c_{12}(n) < 1$, since the direct sound and the reflection superimpose as two signal pairs with different ITD and ILD. Thus, IC can be used as an indicator for whether ITD and ILD cues are similar to free-field cues of sources or not, while ignoring cues related to reflections.

Faller and Merimaa's cue selection method, as the authors point out, can be seen as a "multiple looks" approach for localisation. Multiple looks have been previously proposed to explain monaural detection and discrimination performance with increasing signal duration [Viemeister and Wakefield 1991]. The idea is that the auditory system has a short-term memory of "looks" at the signal, which can be accessed and processed selectively. In the context of localisation, the looks would consist of momentary ITD, ILD, and IC cues. With an overview of a set of recent cues, ITDs and ILDs corresponding to high IC values are adaptively selected and used to build a histogram that provides a statistical description of gathered cues (see Fig. 2.7).

Finally, the binaural processor capitalises on the multiple looks configuration and

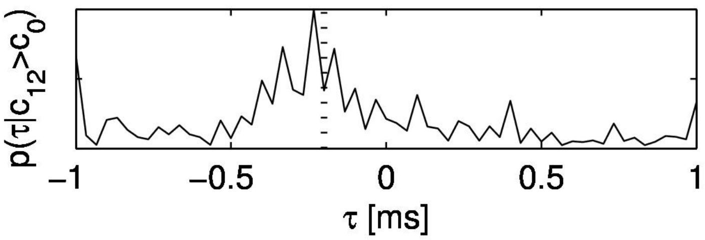


Figure 2.7: Example of the use of an adaptation of the cue selection method proposed by [Faller and Merimaa 2004a] using a 1 s "multiple looks" buffer. Represented in the figure is a histogram of collected ITD cues (τ) corresponding to high IC levels ($c_{12} > c_0$) for a particular frequency channel of a 1 s audio snippet. This histogram is interpreted as a distribution corresponding to the probability of the occurrence of ITD readings, which is then used as a conspicuity map in order to perform a *summary cross-correlogram* over all frequencies (see main text for more details).

implements a simple auditory scene analysis algorithm for detection and extraction of important auditory features to build conspicuity maps and ultimately a saliency map, thus providing a functionality similar to the role of the *external nucleus of the inferior colliculus* (ICx) in the mammalian brain. The first stage of this algorithm deals with figure-ground (i.e. foreground-background) segregation and signal-to-noise ratio. In signal processing, the energy of a discrete-time signal $x(n)$ is given by [Oppenheim and Schafer 1989]

$$E = \sum_{-\infty}^{\infty} |x(n)|^2$$

Using this notion, a simple strategy can be followed to selectively apply the multiple looks approach to a binaural audio signal buffer so that only relevant audio snippets are analysed. This strategy goes as follows: given a binaural signal buffer of N samples represented by the tuple $\{x'_1(n), x'_2(n)\}$, the average of the energies of the component signals $x'_1(n)$ and $x'_2(n)$ is

$$E_{avg} = \frac{\sum_1^N |x'_1(n)|^2 + \sum_1^N |x'_2(n)|^2}{2} \qquad (2.5)$$

and can be used as a noise gate so that only when $E_{avg} > E_0$ ITDs, ILDs and ICs triplets are collected for the buffer, yielding multiple looks values only for relevant signals (just the ITD-ILD pairs corresponding to high IC values are kept in conspicuity maps per frequency channel), while every other buffer instantiation is labelled as irrelevant noise. E_0 can be fixed to a reasonable empirical value or be adaptive, as seems to happen with human hearing. A set of results exemplifying this algorithm is presented on Fig 2.8.

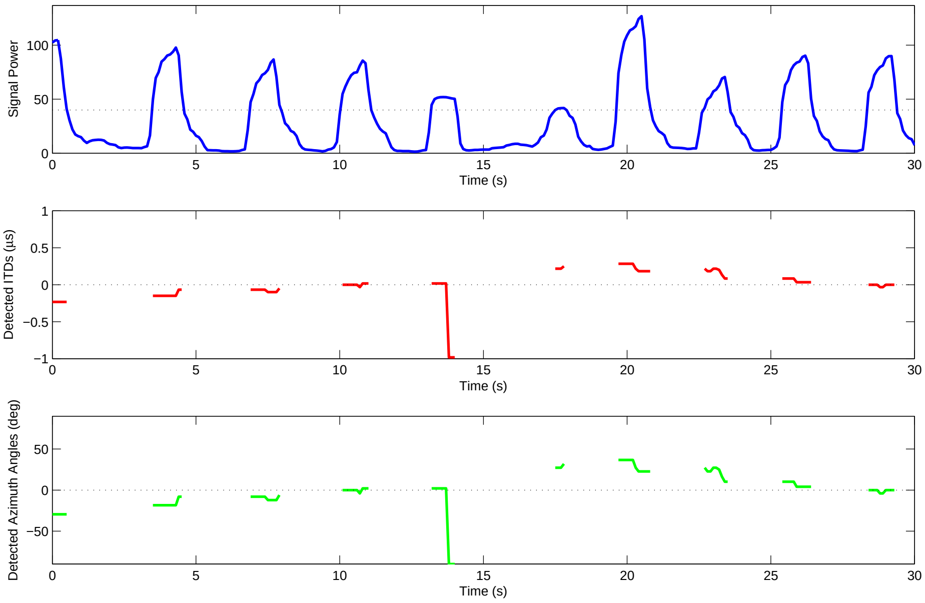


Figure 2.8: Binaural processing results of an approximately 30 second-long audio snippet of a typical "cocktail party" scenario, with the main voice repeatidely calling out "Nicole, look at me" approximately every 3 s, while other voices can be heard coming from sites close to the robotic head, elsewhere in the lab. The active perception head was moved while the main speaker was kept still, first keeping the speaker to the right and slowly travelling towards the centre, then keeping the speaker to the left and again slowly moving towards the centre. Top — the effect of the signal power-based figure-ground segregation noise gate is shown (dashed line represents gate threshold); Middle — ITD estimates for the most salient sound; Bottom — corresponding azimuth estimates. These results show the performance of the binaural processor under difficult conditions, the only "failure" being the estimates corresponding to the 14 s instant: for a signal power above the interest threshold, the background noise (i.e., some other voice in the lab) was more salient than the main voice.

Once the multiple looks information is gathered, since ITDs are proven to be stable across frequencies for a specific sound source at a given azimuth regardless of range or elevation, the ITD conspicuity maps may be summed over all frequencies, in a process similar to what is believed to occur in the ICx, in computational terms known as a *summary cross-correlogram* (again see Fig. 2.7). From the resulting one-dimensional signal, the largest peaks may be taken as having been effected by the most important sound-sources represented in the auditory image. Then, a search is made across each frequency band to find the closest ITD and its ILD pair, for each reference ITD, thus building n-sized vectors (for $m = n - 1$ frequency channels) for each relevant sound source of the form

$$Z = [\tau, \Delta L(f_c^1) \cdots \Delta L(f_c^m)] \tag{2.6}$$

Inertial sensing system

To process the inertial data, we follow the Bayesian model proposed by Laurens and Droulez [2006; 2005] of the human vestibular system, 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.

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.

Considering a small time increment δt the rotation update can be approximated by

$$\boldsymbol{\Theta}^{t+\delta t} = \boldsymbol{\Theta}^t . \boldsymbol{R}(\delta y, \delta p, \delta r) \tag{2.7}$$

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

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

where $\mathrm{c}(\bullet)$ and $\mathrm{s}(\bullet)$ are shorthand notations for $\cos(\bullet)$ and $\sin(\bullet)$, respectively, and the instantaneous angular velocity is defined as the vector:

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

Linear motion of the head is described by the position of the centre of the head in a geocentric reference frame, defined as a position vector $\boldsymbol{M}$. The linear acceleration $\boldsymbol{A}$ is the second derivative of $\boldsymbol{M}$ over time. In our case we are only concerned with the linear acceleration, since gravity will provide an absolute reference for orientation only when $\boldsymbol{A} = 0$. The state of our system at time t is therefore defined by $(\boldsymbol{\Theta}^t, \boldsymbol{\Omega}^t, \boldsymbol{A}^t)$.

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. $\boldsymbol{\Phi}^t = \boldsymbol{\Omega}^t + \eta^t_\Phi$, where η^t_Φ 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. $\boldsymbol{\Upsilon}^t = \boldsymbol{F}^t + \eta^t_\Upsilon$, where η^t_Υ 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 $\boldsymbol{F}$. In a geocentric frame of reference gravity is a vector $\boldsymbol{G} = (0, 0, -9.81)$, and the gravito-inertial acceleration is given by $\boldsymbol{G} - \boldsymbol{A}$, transforming to the egocentric frame of reference we have

$$\boldsymbol{F} = \boldsymbol{\Theta}^{-1}.(\boldsymbol{G} - \boldsymbol{A}) \tag{2.8}$$

The sensor data at time t is therefore defined by $(\boldsymbol{\Phi}^t, \boldsymbol{\Upsilon}^t)$.

As suggested in [Laurens and Droulez 2006], 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 $P(\boldsymbol{\Omega}^t) \propto \mathcal{N}(|\boldsymbol{\Omega}^t|, 0, \sigma_\Omega)$.

2.3 Bayesian Models for Multimodal Perception of 3D Structure and Motion

2.3.1 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;

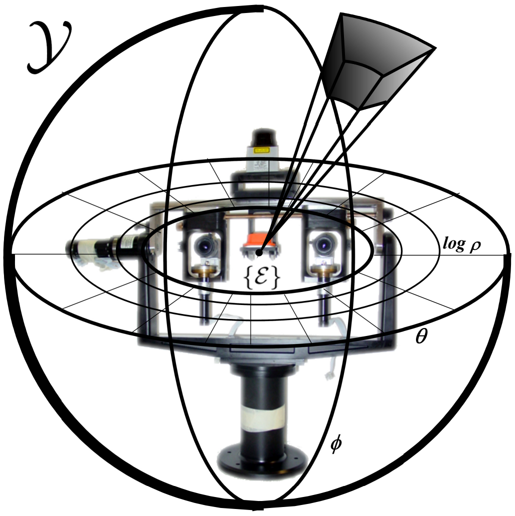


Figure 2.9: Egocentric, log-spherical configuration of the Bayesian Volumetric Maps.

- It is an occupancy grid, allowing for a probabilistic representation of dynamical spatial occupation of the environment, thus encompassing 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. Data association is thus effectively postponed to higher-level processing.

Given these requirements, we chose a *log-spherical* coordinate system spatial configuration (see Figure 2.9) for the occupancy grid that we have developed and will refer to as Bayesian Volumetric Map (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

$$\mathcal{Y} \equiv \left]\log_b \rho_{\text{Min}}; \log_b \rho_{\text{Max}}\right] \times \left]\theta_{\text{Min}}; \theta_{\text{Max}}\right] \times \left]\phi_{\text{Min}}; \phi_{\text{Max}}\right] \tag{2.9}$$

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

$\phi_{\max}$, through:

$$\mathcal{Y} \supset \mathcal{C} \equiv \,]\log_b \rho_{\min}; \log_b \rho_{\max}] \times]\theta_{\min}; \theta_{\max}] \times]\phi_{\min}; \phi_{\max}] \tag{2.10}$$

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 $C = (\log_b \rho_{\max}, \theta_{\max}, \phi_{\max})$.

To compute the probability distributions for the current states of each cell, the *Bayesian Program* (BP) formalism, as first defined by Lebeltel [1999] and later consolidated by Bessière, Laugier, and Siegwart [2008], will be used throughout this text — for more details on this formalism, please refer to Appendix A on page 121.

2.3.2 Multimodal sensor fusion using log-spherical Bayesian Volumetric Maps

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 [Ernst and Bülthoff 2004]: *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* [Ernst and Bülthoff 2004].

We will try to explicitly or implicitly address each of the challenges of sensor fusion as described in [Ernst and Bülthoff 2004] using the BVM, for vision, audition and proprioception (e.g. vestibular sensing). We propose to use proprioception as ancillary information to promote visual and auditory sensing to satisfy the requirements for integration, enumerated above.

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. 2.10 is based on the solution presented by Tay *et al.* [Tay et al. 2007], adapted so as to conform to the BVM egocentric, three-dimensional and log-spherical nature — we will now describe the underlying model in detail. 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 2.3.1. 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 a *hypothetical antecedent* cell of reference cell C. The set of allowed antecedents $\mathcal{A}_C$ of cell C is composed by the $N+1$ cells on the BVM grid from which occupancy of the reference cell at the current time instant is allowed by the model to originate at the previous time instant, caused by the possible movement of an object from one instant to the other, from a specific A_C to C. 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.

- 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 that is considered in the prior distribution for occupancy for cell C if its effective antecedent is assumed to be known, in other words, considering that an object occupying a specific A_C was moved to C.

- V_C denotes the dynamics of the occupancy of cell C as a vector signalling local motion between this cell and 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.

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);

$Z_1, \cdots, Z_S \in \{$"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;

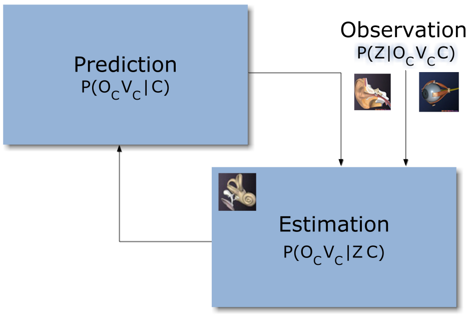
$P(Z_i|V_C\, O_C\, C)$: *direct measurement model* for each sensor i, given by respective sub-BP.

Identification:

None.

Questions:

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



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

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

Figure 2.10: Bayesian Program for the estimation of Bayesian Volumetric Map current cell state (far top), and corresponding Bayesian filter diagram (top middle — it considers only a single measurement Z for simpler reading, with no loss of generality) and respective equation, using two different formulations (bottom middle and far bottom).

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

- $P(A_C)$ is the distribution over all possible antecedents of cell $[C = c]$. In order to represent the fact that cell $[C = c]$ is *a priori* equally reachable from all possible antecedent cells A_C on the map, this probability distribution is chosen to be uniform.

- $P(V_C|A_C)$ is the distribution over all the possible velocities of a certain antecedent of cell $[C = c]$; its parametric form is a histogram.

- $P(C|V_C\,A_C)$ is a distribution that takes into account the probability of $[C = c]$ being reachable from its antecedent $[A_C = a_c]$ with velocity $[V_C = v_c]$. In discrete spaces, this distribution is a Dirac that is equal to one *iff* $c_{\log_b \rho} = a_{\log_b \rho} + v_{\log_b \rho}\delta t$, $c_\theta = a_\theta + v_\theta \delta t$ and $c_\phi = a_\phi + v_\phi \delta t$, thus implying a *constant velocity assumption* for the dynamic model.

- $P(O_C^{-1}|A_C)$ is a conditional distribution that gives the probability of a preceding occupancy state of cell $[C = c]$, given a set of antecedent cells from which state probabilities might propagate. In other words, for each antecedent $[A_c = a_c]$ given the current cell $[C = c]$, this probability is equal to $P(O_{a_c}|a_c)$ taken from the BVM at the preceding time instant.

- $P(O_C|O_C^{-1})$ is the conditional distribution over the current occupancy of cell $[C = c]$ given its preceding occupancy state. If ϵ is the probability of this cell not following the constant velocity hypothesis, it is defined as a transition matrix given by

$$T = \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix}$$

- $P(Z_i|V_C\,O_C\,C)$ is the *direct model* for sensor i. It yields the probability of a measurement given the occupancy O_C and the velocity V_C of cell C. Measurements for all sensors are assumed to have been taken *independently from each other*[3].

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. 2.10. This

[3] This is a relatively safe assumption: in theory, readings from different photoreceptors follow parallel processing pathways, as does audio in respect to the imaging system.

inference effectively leads to the Bayesian filtering formulation as used in the BOF grids. The discrete case of the Bayesian filter can formulated according to the question to be answered through

$$\begin{aligned}
&P(V_C\,O_C\,Z_1\cdots Z_S\,C) = P(V_C\,O_C|Z_1\cdots Z_S\,C)P(C\,Z_1\cdots Z_S)\\
&P(V_C\,O_C|Z_1\cdots Z_S\,C) = \frac{P(V_C\,O_C\,Z_1\cdots Z_S\,C)}{P(C\,Z_1\cdots Z_S)}\\
&P(V_C\,O_C|Z_1\cdots Z_S\,C) =\\
&\qquad\qquad \frac{\sum_{A_C,O_C^{-1}} P(C\,A_C\,O_C\,O_C^{-1}\,V_C\,Z_1\cdots Z_S)}{\sum_{A_C,O_C^{-1},O_C,V_C} P(C\,A_C\,O_C\,O_C^{-1}\,V_C\,Z_1\cdots Z_S)}
\end{aligned} \tag{2.11}$$

Using the decomposition equation given in Fig. 2.10, we also have a more familiar formulation of the Bayesian filter on the same figure (middle), 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}. Finally, substituting (2.11) on this formulation, we get the answer to the Bayesian Program question, the global filtering equation (bottom of Fig. 2.10).

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 $[C = c]$, $\forall k \in \mathbb{N}, 1 \leq k \leq N$,

$$\begin{aligned}
&\alpha_c([O_C = 1], [V_C = v_k]) =\\
&\quad\sum_{A_C,O_C^{-1}} P(A_C)P(v_k|A_C)P(C|v_k\,A_C)P(O_C^{-1}|A_C)P([O_C = 1]|O_C^{-1})
\end{aligned} \tag{2.12a}$$

$$\begin{aligned}
&\alpha_c([O_C = 0], [V_C = v_k]) =\\
&\quad\sum_{A_C,O_C^{-1}} P(A_C)P(v_k|A_C)P(C|v_k\,A_C)P(O_C^{-1}|A_C)P([O_C = 0]|O_C^{-1})
\end{aligned} \tag{2.12b}$$

The prediction step thus consists on performing the computations represented by (2.12) 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, $\forall k \in \mathbb{N}, 1 \leq k \leq N$,

$$\beta_c([O_C = 1], [V_C = v_k]) = \prod_{i=1}^{S} \big(P(Z_i | v_k \, [O_C = 1] \, C)\big) \; \alpha_c([O_C = 1], v_k) \tag{2.13a}$$

$$\beta_c([O_C = 0], [V_C = v_k]) = \prod_{i=1}^{S} \big(P(Z_i | v_k \, [O_C = 0] \, C)\big) \; \alpha_c([O_C = 0], v_k) \tag{2.13b}$$

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, $\forall k \in \mathbb{N}, 1 \leq k \leq N$,

$$l_c(v_k) = \beta_c([O_C = 1], [V_C = v_k]) + \beta_c([O_C = 0], [V_C = v_k]) \tag{2.14}$$

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

$$P(O_C \, [V_C = v_k] | Z_1 \cdots Z_S \, C) = \frac{\beta_c(O_C, [V_C = v_k])}{\sum\limits_{V_C} l_c(V_C)} \tag{2.15}$$

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

$$P(O_C | Z_1 \cdots Z_S \, C) = \sum_{V_C} P(V_C \, O_C | Z_1 \cdots Z_S \, C) \tag{2.16a}$$

$$P(V_C | Z_1 \cdots Z_S \, C) = \sum_{O_C} P(V_C \, O_C | Z_1 \cdots Z_S \, C) \tag{2.16b}$$

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.

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, integrated with other types of proprioception such as what might be emulated by the motor encoders of a robotic active head, will yield measurements of angular velocity and position, that 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 $C = (\log_b \rho_{max}, \theta_{max} - \theta_{inertial}, \phi_{max} - \phi_{inertial}) \in \mathcal{Y}$.*

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

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 [Spence and Squire 2003], 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 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.

Preliminary tests using the BVM update model showed that this, in fact, promotes an effect similar to the well-known temporal ventriloquism (also known as *auditory capture* of visual perception), given the inherently higher auditory measurement frequency as opposed to vision. Spatial ventriloquism (i.e. *visual capture* of auditory perception), 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 perfectly feasible, since inertial readings are available at a much faster rate than visuoauditory perception.

2.3.3 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.

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 [2007]. 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).

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:

$$P(Z\,C\,O_C\,G_C) = \\ P(C)P(O_C|C)P(G_C|O_C\,C)P(Z|G_C\,O_C\,C) \tag{2.17}$$

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. 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 [Yguel et al. 2007].

- $P(Z|G_C\,O_C\,C)$ 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 presented generically in the description of the vision system can be finally formalised as

$$P_k(Z) = L_k(Z, \mu_\rho(k), \sigma_\rho(k)), \begin{cases} \mu_\rho(k) & = \hat{\rho}(\hat{\delta}) \\ \sigma_\rho(k) & = \frac{1}{\lambda}\sigma_{min} \end{cases} \tag{2.18}$$

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 (see Equation (4.6) 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 [Pearl 1988] concerning the relation between vision sensor measurements denoted generically by Z and the corresponding readings δ and λ, described by the calibrated expected value $\hat{\rho}(\hat{\delta})$ and standard deviation $\sigma_\rho(\lambda)$ for each sensor.

Equation (2.18) 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 [Treue, Hol, and Rauber 2000, Born and Bradley 2005, Knill and Pouget 2004, Pouget, Dayan, and Zemel 2000].

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

$$P_k([Z = z]) = \begin{cases} \int_{]-\infty;0]} \mathcal{N}(\mu(k-0.5), \sigma(\sigma_\rho))(u)du, & z \in [0;1] \\ \int_{\lceil z \rceil - 1}^{\lceil z \rceil} \mathcal{N}(\mu(k-0.5), \sigma(\sigma_\rho))(u)du, & z \in]1;N] \\ \int_{]N;+\infty]} \mathcal{N}(\mu(k-0.5), \sigma(\sigma_\rho))(u)du, & z = \text{“No Detection”} \end{cases} \tag{2.19}$$

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 $]k-1;k]$ (see Fig. 2.15 in the Results section for simulation examples).

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 (2.17); 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 (results of inference simulations are presented in Fig. 2.15 in the Results section).

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

Definition 1. $T_c^k \in \mathcal{G}_C$ 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. $\forall(i,j), i \neq j, T_c^i \bigcap T_c^j = \emptyset$

Property 1.2. $\bigcup T_c^k = \mathcal{G}_c \setminus \mathcal{G}_\emptyset$, 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,* $P(T_c^k) = P_{\text{Empty}}^{k-1}(1 - P_{\text{Empty}})$.

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,* $P(T_c^k) = P_{\text{Empty}}^{k-2}(1 - P_{\text{Empty}})$.

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 (see Fig. 2.14 in the Results section for corresponding simulation results).

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 [Yguel et al. 2007],

$$\begin{aligned}
&\forall Z \neq \text{“No Detection”},\\
&P(Z|[O_C = 1]\,C) =\\
&\quad = \sum_{g_c \in \mathcal{G}_C} P([G_c = g_c])P(Z|[O_C = 1]\,[G_c = g_c]\,C)\\
&\quad = \sum_{k=1}^{c-1} P(T_c^k)P_k(Z) + (1 - \sum_{k=1}^{c-1} P(T_c^k))P_c(Z)\\
&\quad = \sum_{k=1}^{c-1} P_{\text{Empty}}^{k-1}(1 - P_{\text{Empty}})P_k(Z) + P_{\text{Empty}}^{c-1}P_c(Z) \qquad (2.20)
\end{aligned}$$

Equation (2.20) 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,

$$\begin{aligned}
&Z = \text{“No Detection”},\\
&P(Z|[O_C = 1]\,C) =\\
&\qquad = 1 - \sum_{r \neq \text{“No Det.”}} P([Z = r]|[O_C = 1]\,C) \qquad (2.21)
\end{aligned}$$

$$
\text{Program}\begin{cases}
\text{Description}\begin{cases}
\text{Specification}\begin{cases}
\text{Relevant variables:}\\
C\text{: cell identifier,}\\
\quad\text{stored as a 3-tuple of cell coordinates } (\log_b \rho_C, \theta, \phi);\\
Z \in \{\text{“No Detection”}\} \cup \mathcal{Z}_{\text{VisDepth}}\text{: sensor depth measurement along line-of-sight } (\theta, \phi);\\
O_C\text{: binary value describing the occupancy of cell } C;\\
G_C \in \mathcal{G}_C \equiv \mathcal{O}^{N-1}\text{: state of all cells in the line-of-sight except for } C.\\
\text{Decomposition:}\\
P(Z\, C\, O_C\, G_C) =\\
\quad P(C)P(O_C|C) \cdot \underbrace{P(G_C|O_C\, C)P(Z|G_C\, O_C\, C)}_{\text{Gives } P(Z|O_C\, C) \text{ through } \sum_{G_C}.}\\
\text{Parametric forms:}\\
P(C)\text{: uniform;}\\
P(O_C|C)\text{: uniform or prior estimate;}\\
P(G_C|O_C\, C)\text{: unknown, apart from dependency on number of occupied cells;}\\
P(Z|G_C\, O_C\, C)\text{: probability of a measurement by sensor,}\\
\quad\text{knowing first occupied cell is } [C = k] \equiv \textit{elementary sensor model } P_k(Z)\text{, equation (2.19).}
\end{cases}\\
\text{Identification:}\\
\text{Calibration for } P_k(Z) \Rightarrow P(Z|G_C\, O_C\, C).
\end{cases}\\
\text{Question (given cell velocity } v_c\text{):}\\
P(Z|v_c\, o_c\, c) \equiv P(Z|o_c\, c)
\end{cases}
$$

Figure 2.11: Bayesian Program for vision sensor model of occupancy.

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*),

$$
\begin{aligned}
&\forall Z \neq \text{“No Detection”},\\
&P(Z|[O_C = 0]\, C) =\\
&\quad = \sum_{g_c \in \mathcal{G}_C} P([G_c = g_c])P(Z|[O_C = 0]\, [G_c = g_c]\, C)\\
&\quad = \sum_{k=1, k\neq c}^{N} P(T_c^k)P_k(Z) + P(\mathcal{G}_\emptyset)\delta_{Z=\text{“No Detection”}}\\
&\quad = \sum_{k=1}^{c-1} P_{\text{Empty}}^{k-1}(1 - P_{\text{Empty}})P_k(Z) +\\
&\quad + \sum_{k=c+1}^{N} P_{\text{Empty}}^{k-2}(1 - P_{\text{Empty}})P_k(Z) + P_{\text{Empty}}^{N-1}\delta_{Z=\text{“No Det.”}}
\end{aligned}
\tag{2.22}
$$

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 $\mathcal{G}_\emptyset$,

$$
\begin{aligned}
&Z = \text{“No Detection”}, \\
&P(Z|[O_C = 0]\, C) = \\
&\quad = 1 - (\sum_{r}^{N} P([Z = r]|[O_C = 0]\, C)) + P_{\text{Empty}}^{N-1} \delta_{Z=\text{“No Det.”}}
\end{aligned}
\tag{2.23}
$$

The Bayesian Program that summarises this model is presented on Fig. 2.11.

Audition sensor model

The direct audition sensor model is formulated as the first question of the Bayesian Program in Fig. 2.12, 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 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, 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.

Vestibular sensor model

At time t the Bayesian program of Fig. 2.13 computes the probability distribution of the current state $\boldsymbol{\xi}^t$ given all the previous sensory inputs the initial distribution $\boldsymbol{\xi}^t$ — to simplify notation, state variables are grouped in a vector $\boldsymbol{\xi}^t = (\boldsymbol{\Theta}^t, \boldsymbol{\Omega}^t, \boldsymbol{A}^t)$ and sensor variables in a vector $\boldsymbol{S}^t = (\boldsymbol{\Phi}^t, \boldsymbol{\Upsilon}^t)$. The inference of current state is done by applying the conjunction and marginalisation rule, applying a summation over state variables at the previous time step so that no decision is taken about these values,

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 $[\tau, \Delta L(f_c^1) \cdots \Delta L(f_c^m)]$;

(see equation (2.6): $\tau \equiv$ 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:

$P(Z\, C\, S_C\, O_C) =$

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

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, empirically chosen or learned from scene statistics;

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

$P(\tau|S_C\, O_C\, \theta_{\max}) \equiv P(\tau|S_C\, \theta_{\max})$: 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 2.12: Bayesian Program for binaural sensor model. On the right 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 **from this information alone** if it is in this condition due to a sonorous object or not (right column).

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:

$P(\boldsymbol{\xi}^t\, \boldsymbol{\xi}^{t-\delta t}\, \boldsymbol{S}^t ... \boldsymbol{S}^0) =$

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

$.P(\boldsymbol{\Omega}^t).P(\boldsymbol{A}^t).P(\boldsymbol{\Theta}^t|\boldsymbol{\Theta}^{t-\delta t}\, \boldsymbol{\Omega}^t)$

$.P(\boldsymbol{\xi}^{t-\delta t}\, \boldsymbol{S}^{t-\delta t} ... \boldsymbol{S}^0)$

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;

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

Identification:

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

Question:

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

Figure 2.13: Bayesian Program for processing of inertial data.

summarising all the past in the answer to the estimation question in the previous time step, and can be formulated by:

$$P(\boldsymbol{\xi}^t|\boldsymbol{S}^t ... \boldsymbol{S}^0) = \frac{1}{K}\sum_{\xi^{t-\delta t}} P(\boldsymbol{S}^t|\boldsymbol{\xi}^t)P(\boldsymbol{\Omega}^t)P(\boldsymbol{A}^t)P(\boldsymbol{\Theta}^t|\boldsymbol{\Theta}^{t-\delta t}\, \boldsymbol{\Omega}^t)P(\boldsymbol{\xi}^{t-\delta t}\, \boldsymbol{S}^{t-\delta t} ... \boldsymbol{S}^0)$$

where:

- K is a normalisation constant;

- $P(\boldsymbol{S}^t|\boldsymbol{\xi}^t) = P(\boldsymbol{\Phi}^t|\boldsymbol{\Omega}^t)P(\boldsymbol{\Upsilon}^t|\boldsymbol{F}^t)P(\boldsymbol{F}^t|\boldsymbol{\Theta}^t\, \boldsymbol{A}^t)$ is the sensor model, i.e., the probability distribution of sensor inputs given the sate. $P(\boldsymbol{\Phi}^t|\boldsymbol{\Omega}^t)$ and $P(\boldsymbol{\Upsilon}^t|\boldsymbol{F}^t)$ are Gaussians and $P(\boldsymbol{F}^t|\boldsymbol{\Theta}^t\, \boldsymbol{A}^t)$ a dirac, equal to 1 if and only if equation 2.8 is verified;

- $P(\boldsymbol{\Omega}^t)$, $P(\boldsymbol{A}^t)$ represent *a priori* knowledge about state variables, both Gaussians;

- $P(\boldsymbol{\Theta}^t|\boldsymbol{\Theta}^{t-\delta t}\, \boldsymbol{\Omega}^t)$ is the system dynamic model for state variable $\boldsymbol{\Theta}$, i.e., the probability distribution of rotation $\boldsymbol{\Theta}$ given the previous rotation and current angular velocity. $P(\boldsymbol{\Theta}^t|\boldsymbol{\Theta}^{t-\delta t}\,\boldsymbol{\Omega}^t)$ is a dirac, equal to 1 if and only if equation 2.7 is verified;

- $P(\boldsymbol{\xi}^{t-\delta t}\,\boldsymbol{S}^{t-\delta t}...\,\boldsymbol{S}^0)$ is the probability distribution computed at the last time step, i.e., from the previous iteration of the Bayesian filter.

We can see also that the first-order Markov assumption is present in both the state dynamic model and sensor model: time dependence has a depth of one time step. The stationarity assumption is also implicit: models do not change with time. The filter iterates for each new time step, but the relationships between these variables remain the same for all time steps. This greatly reduces the complexity.

For the implementation the space of $\xi^{t-\delta t}$ that needs to be scanned has 3 dimensions: $\boldsymbol{\Theta}^{t-\delta t}$. For a given $\xi^{t-\delta t}$, the space of possible ξ^t has 3 dimensions, so the total search space has 6 dimensions. Following the Bayesian model implementation proposed by Laurens and Droulez [2006], we used a Particle Filter to perform the inference. A set of N particles, $\xi^{i,t}$, sample the state search space, and each one has an associated weight $w^{i,t} = P(\xi^{i,t})$. Starting from $\xi^{i,t-\delta t}$, we draw values for the Gaussians and apply equations (2.7) and (2.8) to obtain $\xi^{i,t}$. The weighing factor is updated to $w^{i,t} = w^{i,t-\delta t}.P(\xi^{i,t})$. Resampling is applied, so that unlikely particles are deleted and likely ones are duplicated, in order to avoid having all particles drift towards improbable states. At each iteration a new set of N samples is drawn from the previous set of particles. Each particle of the previous set has a probability w^i to be chosen for each new particle. The weights in the new set are levelled to $1/N$.

2.4 Results and Conclusions

In this section, results concerning the models described herewith, yielded both from simulation and from preacquired experimental sensor data, are presented, and conclusions are drawn.

More specifically, on Figs. 2.14 and 2.15, results for simulations using the direct vision sensor model to process an idealised projection line extending from a pixel on the cyclopean view knowing the first occupied cell on the BVM along that line, and inference applying the vision sensor model on synthetic visual observations, respectively, are shown. Effects of visual occlusion can clearly be seen: for a given depth estimate $\hat{\rho}$, inference assumes all cells much closer to the observer than the estimate to be most probably empty, while no relevant evidence regarding the probability of occupancy is

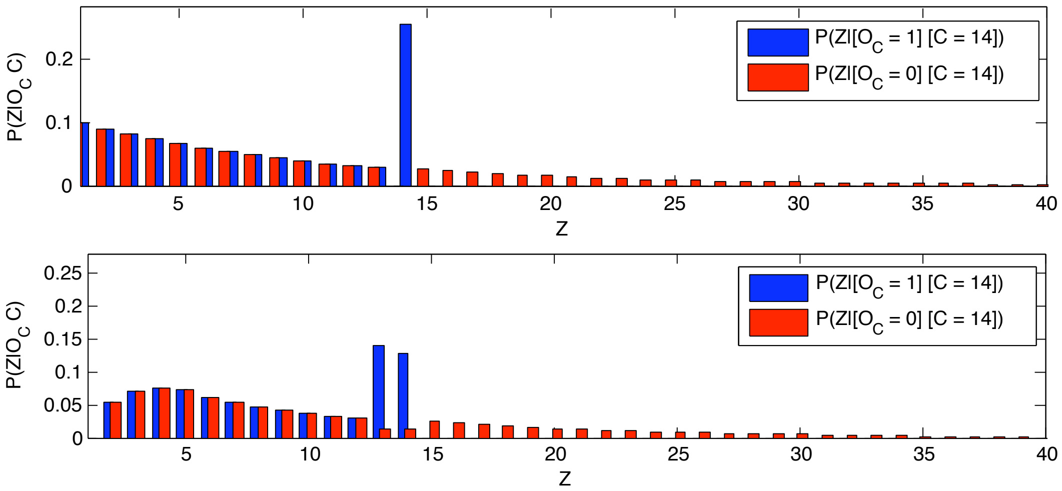


Figure 2.14: Simulation results for direct vision sensor model for $[C = 14]$, given $P_{Empty} = .9$, $N = 40$, $\rho_{Min} = 1000\,\text{mm}$ and $\rho_{Max} = 11000\,\text{mm}$, considering both occupied and unoccupied states. Top: ideal sensor model (Dirac). Bottom: Gaussian elementary sensor model with $\sigma_\rho = 1\,\text{mm}$. Note that for the ideal sensor model, precision is maximal and aggregation is complete at $P([Z = 14] | [O_C = 1][C = 14])$; additionally, note that for either of the presented cases, for $Z << 14$, $P(Z|[O_C = 1][C = 14]) = P(Z|[O_C = 0][C = 14])$, for $Z = 14$, $P(Z|[O_C = 1][C = 14]) >> P(Z|[O_C = 0][C = 14])$, and for $Z >> 14$, $P(Z|[O_C = 1][C = 14]) \approx 0$, while $P(Z|[O_C = 0][C = 14]) > 0$. This reflects the assumption coded in the model that, *when C is known to be occupied* (i.e. $[O_C = 1]$), cells farther from the origin than $[C = 14]$ are occluded, and hence do not yield visual readings.

assumed to be collected for cells much farther than the estimate, and therefore prior knowledge gathered from previous inference steps becomes preponderant for these cells.

On Figs. 2.16 and 2.17, on the other hand, two examples of results of a single step of inference (i.e. no local motion is estimated, since all velocities are equally probable at startup) using the binaural sensor model on the BVM are presented. Full 3D auditory localisation has rarely been explored in robotic applications (see, for example, [Calamia 1998] for a review on this subject); as can be seen in these results, this work contributes with a novel probabilistic solution that produces these localisation estimates based on binaural cues alone yielded by a robust binaural processing unit.

Results for the Bayesian vestibular sensor model are presented on Fig 2.18. For comparison with our Bayesian implementation, the Xsens IMU firmware MotionTracker was used to provide attitude estimation. The Motion Tracker implements a weighed filtering of the accelerometer, gyro and magnetic data to provide sensor angular position, including an adaptive filter used to correct for magnetic disturbances. The added data from the magnetic sensor enables the firmware estimation filter to provide a relative

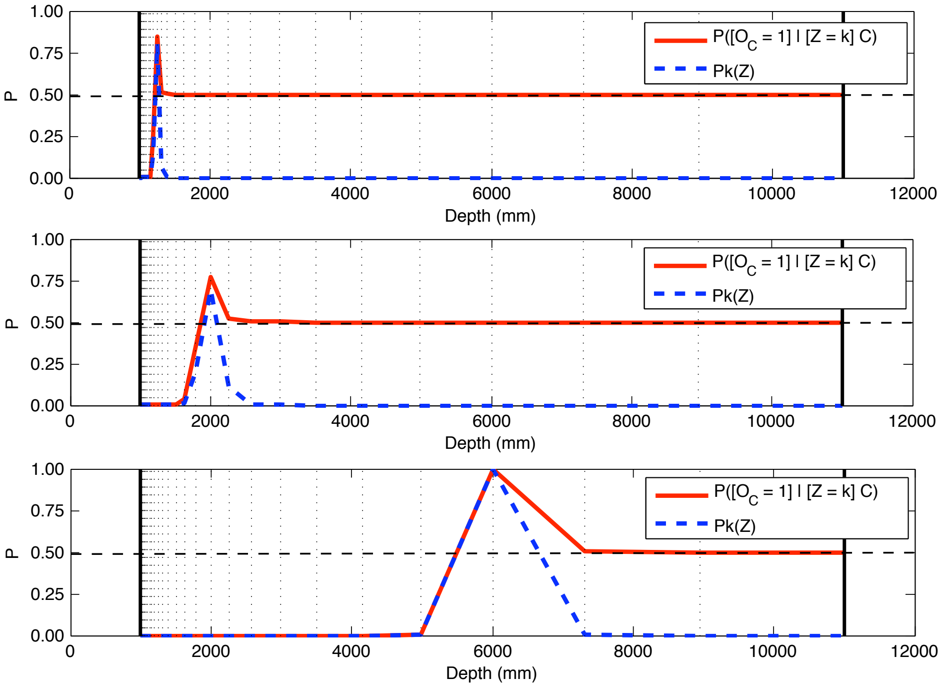


Figure 2.15: Simulation results of inference using vision sensor model. For all cases, $N = 40$, and $\rho_{Min} = 1000\,\mathrm{mm}$ and $\rho_{Max} = 11000\,\mathrm{mm}$ (delimited by full vertical lines), which results in $b \approx 1.2589\,\mathrm{mm}$; each cell C is delimited by black, dashed vertical lines. In any of the graphs, the full red traces correspond to the result of inference (the horizontal axis in this case represents positions in depth throughout the line-of-sight and k the vision sensor measurement) and the full blue traces correspond to the Gaussian elementary sensor models (the horizontal axis in this case represents depth readings from the vision sensor and k the only occupied cell in the line-of-sight). Top: results for $\sigma_\rho = 20\,\mathrm{mm}$, with $b^k + \rho_{Min} = 1200\,\mathrm{mm}$. Middle and bottom: results for $\sigma_\rho = 100\,\mathrm{mm}$, with $b^k + \rho_{Min} = 2000\,\mathrm{mm}$ for the former and $b^k + \rho_{Min} = 5000\,\mathrm{mm}$ for the latter. To note: the fact that Bayesian inference correctly yields the effects described originally by Elfes [1989; 1992], and the effects of the logarithmic partitioning of depth and of the soft evidence conveyed by the elementary sensor model.

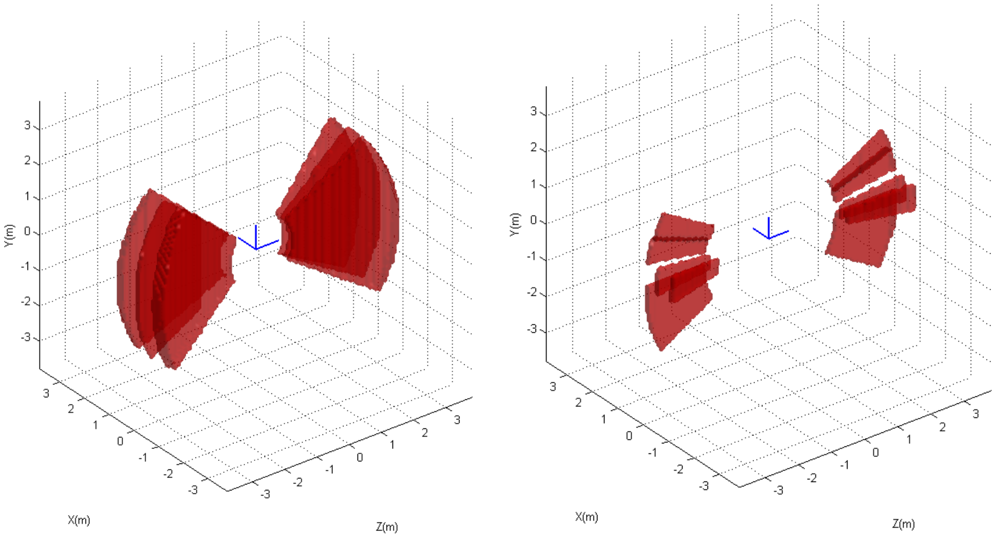


Figure 2.16: Inference results for the processing of an audio snippet of a human speaker placed in front of the binaural perception system — these results correspond to a single inference step using the BVM Bayesian filter. Cells within the log-spherical sensor-space with probabilities of occupancy greater than .75 are depicted in red, and the egocentric referential in blue (X-axis, Y-axis and Z-axis indicate right-to-left, upward and forward directions, respectively). On the left, result of inference using ITDs only; on the right, result of adding ILDs: note the effects on distance and elevation.

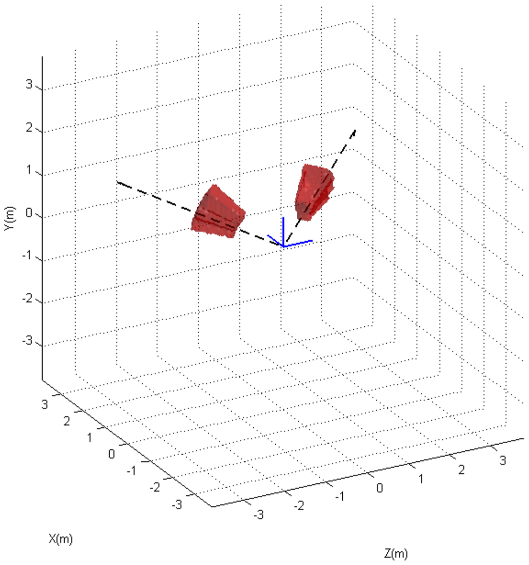


Figure 2.17: Inference results for the processing of an audio snippet of a sound-source placed at $\rho = 1320\,\mathrm{mm}$, $\theta = 36^o$, $\phi = 20^o$. All that is depicted has the same meaning as in Fig. 2.16; two dashed directional lines at (θ, ϕ) and $(180^o - \theta, \phi)$ have been additionally plotted to demonstrate the effect of front-to-back confusion. This phenomenon can be countered in two different ways: either by rotating the perceptual system or by using artificial *pinnae*. The fact that $\theta >> 0^o$ means that precision in elevation and distance is improved as compared to Fig. 2.16.

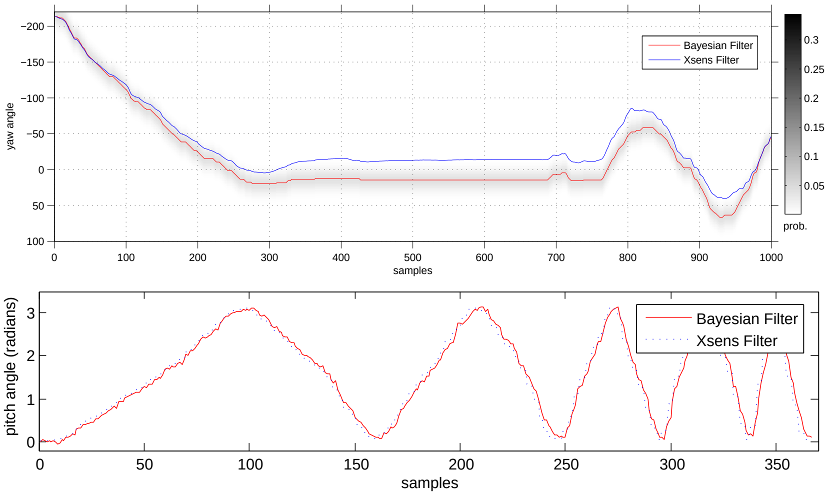


Figure 2.18: Results of Bayesian processing of inertial data. Top: result for yaw angle at 100 Hz sample rate. The magnetic data enables the Xsens filter to slightly outperform the Bayesian filter, as seen in the plot the probabilistic value accumulates some drift. Bottom: result for pitch where gravity is taken into account to bound accumulated drift. Here a particle filter with just 200 particles was used, using $\sigma_A = 0.3\,{}^{\mathrm{m}}/_{\mathrm{s}^2}$ and $\sigma_\Omega = 1\,{}^{\mathrm{rad}}/_{\mathrm{s}^1}$.

ground truth for our experimental work. These results show that the Bayesian model satisfactorily reproduces human vestibular perception using inertial sensors.

Finally, results of multimodal perception of 3D structure and motion using the BVM, yielded after several steps of inference using the Bayesian filter, are presented on Fig. 2.19 — the set of possible local motion cases used in this example and all other experiments presented in this dissertation correspond to the set of antecedents corresponding to the 6 adjacent cells A_C aligned with the 3 spherical directions (i.e. distance, azimuth and elevation) and the cell of reference C itself (i.e. corresponding to absence of motion). Objects within the horopter are reasonably detected, showcasing the advantages of the BVM-BOF framework in terms of the goals presented on the introductory section. Moreover, belief over 3D structure and motion of the objects of the environment and the observers own (rotational) egomotion are conveniently represented using the BVM framework, which has been conceptually designed to be the most appropriate solution for active perception. This can be fully realised in the following chapter.

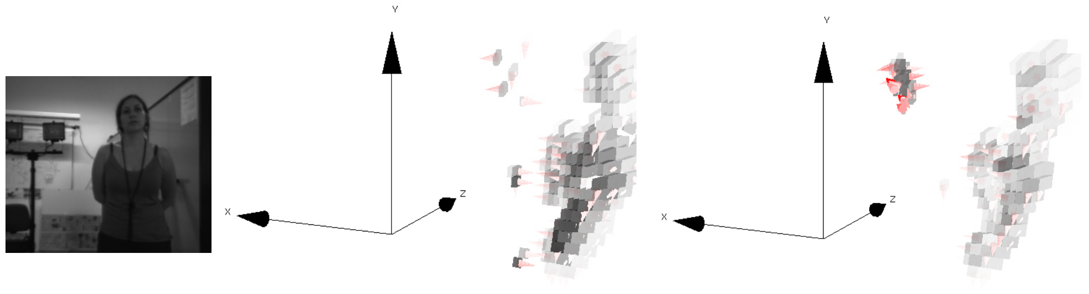


Figure 2.19: Results of multimodal perception of 3D structure and motion using the BVM. A scene consisting of a female speaker talking in a cluttered lab, where several other people (outside the visual field of the IMPEP) were working as usual, thus promoting a typical "cocktail party" effect, is observed by IMPEP V.1 (the scene is shown on the left) and afterwards processed offline by the BVM Bayesian filter. Centre: result of 20 steps of filtering; right: result of 22 steps of filtering, corresponding to the instant where the spotlight on the left of the scene was moved further to the left (in a non-controlled fashion). These results show a projection of the log-spherical configuration onto Euclidean space; the parameters for the BVM are as follows: $N = 40$, $\rho_{Min} = 1000\,\text{mm}$ and $\rho_{Max} = 11000\,\text{mm}$, $\theta \in [-180^o, 180^o]$, with $\Delta\theta = 1^o$, and $\phi \in [-90^o, 90^o]$, with $\Delta\phi = 2^o$, corresponding to $40 \times 360 \times 90 = 1,296,000$ cells. All BVM cells with probability of occupancy greater than 60% are painted using an average of grayscale levels taken from the cyclopean view, with the respective alpha-channels manipulated so as to obtain a degree of transparency proportional to uncertainty (i.e. 60% probability corresponds to maximum transparency, while 100% probability corresponds to maximum opacity); of these cells, all which have non-uniform probability distributions for local motion are associated to velocity vector fields depicted in red (vectors depict directions corresponding to velocities with highest probability), following the same transparency rule as occupancy. In both BVMs, two of the closest objects to the IMPEP are visible in the results: the speaker and the spotlight. As expected, cells corresponding to the moving spotlight on the second BVM exhibit high probability of local motion — since there seems to be some approximation of the spotlight to the perception system during its translation, some local motion directions are consistent with an expansion effect.

Chapter 3

Baseline Research on Multimodal Motion Perception

3.1 Introduction

Historically, the majority of the research on multimodal perception has focussed on interactions in the perception of stationary stimuli, but given that the majority of stimuli in the world move, an important question concerns the extent to which principles derived from stationary stimuli also apply to moving stimuli. A key finding emerging from recent work with moving stimuli is that our perception of stimulus movement in one modality is frequently, and unavoidably, modulated by the concurrent movement of stimuli in other sensory modalities. Visual motion has a particularly strong influence on the perception of auditory and tactile motion [Soto-Faraco et al. 2004].

Given such extensive interactions between the senses in the perception of motion, it is interesting to speculate on their neural bases. According to the traditional view of multisensory integration, information regarding motion in each sense is initially processed independently in modality-specific (or unimodal) brain areas. It is only at later stages of processing that information from different modalities converges in higher-order association areas. In accordance with this idea visual motion processing has repeatedly been shown to involve visual area V5/MT+. Moreover, lesions in this part of the brain appear to impair just visual motion processing, while leaving auditory and tactile motion processing intact. Similarly, researchers have also demonstrated the selective involvement of certain areas (such as the *planum temporale*, the *inferior and superior parietal cortices*, and the *right insula*) in the processing of auditory motion, whereas the *primary and secondary somatosensory areas* (SI and SII), located in the *postcentral gyrus*, play a major role in the perception of tactile motion [Soto-Faraco et al. 2004].

Recent studies have demonstrated that, in addition to these putatively modality-specific motion-processing areas, there are a number of brain areas that appear to be responsive to motion signals in more than one sensory modality. Using functional magnetic resonance imaging (fMRI), these studies have shown that areas of the *intraparietal sulcus*, as well as the *precentral gyrus* (*ventral premotor cortex*, or PMv), can be activated by auditory, visual, or tactile motion signals. These findings converge with the results of animal studies using single-cell recording techniques, as such studies have found neurons that are responsive to both visual and somatosensory motion. Moreover, these neurons are found in brain regions that are homologous to the regions in the human brain that neuroimaging studies have indicated are responsive to multisensory motion signals. It should be noted, though, that these recent fMRI studies failed to control for certain variables, such as non-motion-related activation (i.e., the baselines used were very different across the different modalities) and attentional factors (i.e., moving stimuli might simply capture attention more than stationary stimuli, and this might account for the differences in neural activity reported). However, when taken together, the evidence increasingly supports the idea that a network of brain areas is critically involved in processing motion information from more than one sensory modality [Soto-Faraco et al. 2004].

The challenge for the future will be to develop novel experimental paradigms that can integrate behavioural and neuroscientific approaches in order to refine our understanding of multisensory contributions to the perception of movement (for an in-depth review on this particular subject, refer to Soto-Faraco, Spence, Lloyd, and Kingstone [2004]).

3.2 Baseline Studies of Human Visuoauditory Motion Perception

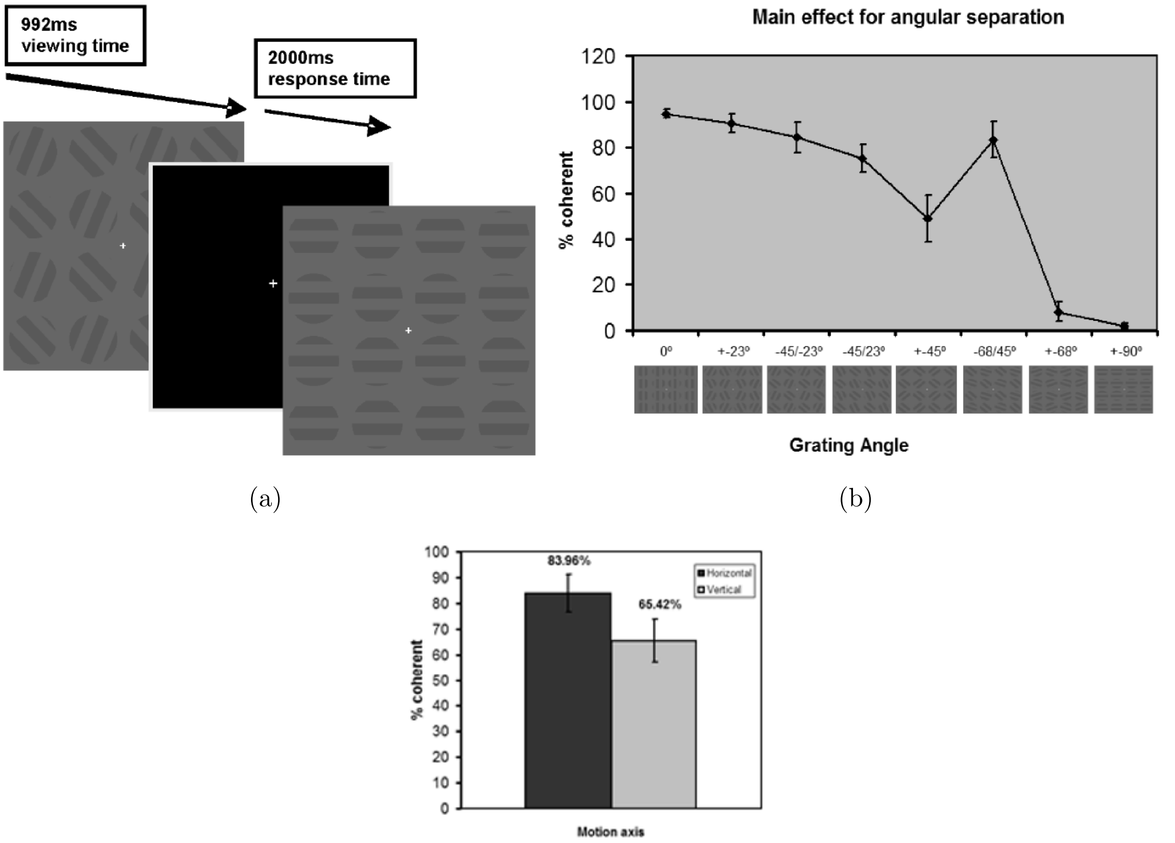
3.2.1 Coherence of visual motion cues is stronger along the horizontal than the vertical meridian: a new light on the ecology of human vision

The nature of ecological constraints on motion integration remains largely unexplored in which concerns perceptual decision rules. The most interesting features of the mammalian world lie either along the line of horizon or below it, leading to differences in psychophysical performance when stimuli are presented across vertical/horizontal meridians [Previc 1990, Rubin et al. 1996, Carrasco et al. 2001, Liu et al. 2006, Silva et al. 2008]. However, deciding on the real nature of ambiguous dynamic stimuli may be crucial for survival and the relevant rules may well go beyond psychophysical asymmetries in visual performance.

Here we chose to investigate horizontal-vertical anisotropies on perceptual grouping of visual dynamic cues. These were either discontinuous moving gratings seen through multiple spatially distributed apertures [Alais et al. 1998] (Fig. 3.1, Experiment 1 and 2) or 2 continuous gratings (plaid stimuli) overlapping in a single aperture [Castelo-Branco et al. 2000; 2002, Castelo-Branco et al. 2006, Schmidt et al. 2006, Kozak and Castelo-Branco 2008] (Fig. 3.2). In the former case, stimuli could be seen either as containing one coherent motion signal or several component incoherent motion signals. When motion integration occurred it was based on independent, distributed local motion signals. In the latter case, requiring integration of contiguous contours, either a single coherent direction could be perceived or, alternatively, the direction of the two component gratings (incoherent motion).

The first two experiments required integration of discrete moving contours across space, essentially mimicking the situation encountered by the visual system when contours of moving objects are discontinuous and partially occluded. We have compared coherence of stimuli moving either to the left or to the right with upward/downward moving stimuli. We hypothesised that if known ecological constraints are relevant, subjects should perceive more coherent motion for grating stimuli moving horizontally than vertically and (if visual experience is also determinant) to the right than for their equivalents moving to the left. In the third experiment, whereby perceptual coherence was achieved by integration of spatially contiguous cues, predictions were identical.

In the first two experiments, (see Fig. 3.1(a), $n = 11$ subjects) grating motions could be directed along multiple global visual movement directions ("left", "right", "up" and "down"). The local grating orientations were either symmetrical ($\pm 0^o, \pm 23^o, \pm 45^o, \pm 68^o$ and $\pm 90^o$ with respect to the vertical or horizontal axis) or asymmetrical ($-23^o/+45^o, -68^o/-45^o$ and $-45^o/-23^o$ with respect to the vertical or horizontal axis), thus providing a total of 8 different local orientation/motion conditions. Experiment 1 used left, right and upward conditions and Experiment 2 was identical except that it used all cardinal directions. These experiments tested subjects' ability to group the independent motions of two gratings seen through 16 apertures into a single coherent motion while varying the overall motion direction of the gratings



(a)

(b)

(c)

Figure 3.1: Description of baseline Experiments 1 and 2 of visual motion perception. (a) Illustration of sequence and timing of stimulus presentation. The array of apertures could either be perceived as independently moving gratings or as one globally moving pattern. (b) Global motion coherence percepts as a function of local grating orientation. (c) Bar graphs displaying the percent time of reportedly perceived coherence for motion along the vertical and horizontal axes.

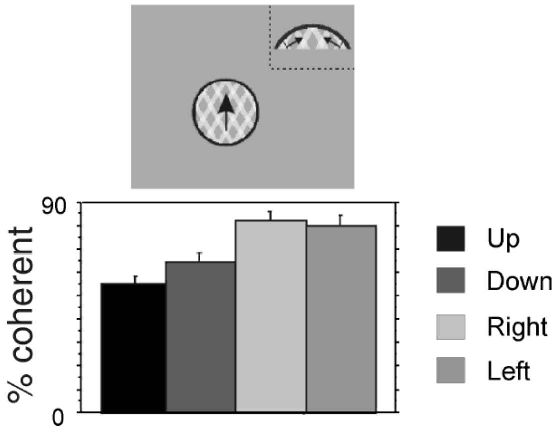


Figure 3.2: Description of baseline Experiment 3 of visual motion perception. Overlapped local motion cues (top inset) cause vertical-horizontal motion coherence asymmetries, in contrast to right-left and up down asymmetries.

(left, right, and upward motion) and the angular separation between the gratings (grating angles $0^o, \pm 23^o, \pm 45^o, \pm 68^o$ and $\pm 90^o, -23^o/+45^o, -68^o/-45^o$ and $-45^o/-23^o$). The percentage of coherent responses for each condition was recorded and data were analyzed using a GLM repeated measures analysis.

Results indicate a significant main effect of grating angle ($F = 47.250; P < .001$), with global coherence reported more often for small angular separations (Fig. 3.1(b)). We found that the greater the angular separation between the two independent gratings, the less coherence subjects perceived (except the asymmetrical -68/45 condition), thereby replicating the findings by Alais et al. [1998].

On the basis of the percentages coherence for the different grating angles in all subjects, we then selected the most "ambiguous" conditions ($-45/23^o$ and $+-45^o$). This choice was based on the fact that only ambiguous conditions challenge perceptual decision mechanisms, and also to prevent floor/ceiling effects. Overall motion direction to the right was found to be perceived as the most coherent, followed by leftward motion and upward visual motion yielded the least number of key presses for both ambiguous angles. GLM indicated a significant main effect of overall visual motion direction ($F = 11.7; P < .002$). Within-subject contrasts revealed that right and left visual movement direction yielded significantly different coherence percepts ($F = 5.537; P = .04$) with more coherence reported for rightward than leftward moving stimuli. However the more robust effects were found when contrasting horizontal directions with upward motion coherence. Accordingly, and concerning right and left *vs* up visual movement directions we found significant and robust direction effects ($F = 14.293; P = .004$ and $F = 9.975; P = .01$, respectively).

Experiment 2 confirmed a significant main effect for grating angle on stimulus ambiguity ($F = 19.692, P << .001$). After pooling data of the two ambiguous angles ($-45/23^o$ and $\pm 45^o$), a repeated-measures analysis replicated also a significant effect of visual movement direction ($F = 6.179; P < .006$). The within-subject contrasts, again, indicate a bias for rightward compared to leftward moving stimuli ($F = 9.275; P = .029$). As expected, Coherence of up and down visual movement directions was lower than for stimuli moving to the right ($F = 22.237; P = .005$ and $F = 11.179, P = .020$, respectively), but not from the leftward moving stimuli ($P = .205; P = .189$). When comparing horizontal (left and right) versus vertical motion directions (up and down), we have found a significant difference ($P < 0.05$) between these axes. Subjects perceived more coherence for stimuli moving along the horizontal axis than along the vertical axis, confirming the horizontal-vertical bias for the grouping of our stimuli (Fig. 3.1(c)).

Fifteen subjects participated in Experiment 3 (testing coherence of contiguous cues), all with normal or corrected-to-normal visual acuity. Observers were asked to give continuous report whether they perceived non-coherence or coherence within a 5° central circular region. Coherent plaid motion directions were randomized across the four cardinal directions. Repeated measures GLM analysis showed a significant main effect of direction of motion ($F = 22.4, p << 0.001, n = 15$, Fig. 3.2). Both vertical axes of motion (upward and downward) showed significantly lower motion coherence than horizontal axes (rightward and downward; $p << 0.001$ for up *vs* right; $p = 0.001$ for up *vs* left; $p = 0.008$ for down *vs* right; $p = 0.001$ for down *vs* left, Bonferroni *post hoc* analyses). Interestingly, left/right asymmetries were not present under contiguity conditions.

3.2.2 Visual direction of bistable coherent motion biases auditory motion perception: evidence for asymmetric dominance of visual *vs* auditory context in perceptual decision

Multisensory perception of motion is of great ecological relevance given the noisy, ambiguous and even incomplete descriptions coming from each sensory modality and the need to achieve an integrated and coherent perception of the environment.

In particular, crossmodal interactions between vision and audition are paramount in their importance concerning fast perceptual assessment of the surroundings and decision. In fact, while vision may provide the most salient information with regard to stimulus motion, audition can also provide important cues, particularly when stimuli are occluded, or else move outside the current field of view, such as when objects move behind the head [Soto-Faraco, Kingstone, and Spence 2003]. Numerous behavioural studies of multisensory interactions have been conducted for several years now, namely on how the presentation of stimuli in one modality (either moving or stationary) affect the perception of motion of stimuli presented in another modality (also either moving or stationary) — see, for example, Bertelson, Vroomen, de Gelder, and Driver [2000], Vroomen and de Gelder [2000], Meyer and Wuerger [2001], Alais and Burr [2004], Meyer, Wuerger, Rohrbein, and Zetzsche [2005], López-Moliner and Soto-Faraco [2007], Zhou, Wong, and Sekuler [2007], Mozolic, Hugenschmidt, and Peiffer [2008], Jain, Sally, and Papathomas [2008]. These studies have addressed multiple specific issues concerning crossmodal interactions, such as spatiotemporal modulation, level of processing, the role of dynamic information (e.g, stimulus onset asynchronies), the role of perceptual grouping, and modality dominance - for a comprehensive review see Soto-Faraco et al. [2003].

As Soto-Faraco et al. point out, an important aspect in the interpretation of perceptual asymmetries is the relative distribution of attention across the different sensory modalities (see also Mozolic et al. [2008]); in fact, several researchers have postulated that visual input may often dominate over input from other modalities because of a pervasive bias to attend preferentially toward the visual modality, rather than any of the remaining modalities [Soto-Faraco et al. 2003]. Even within modality, perceptual and neural responses reflect the way featural attention is distributed [Castelo-Branco et al. 2007].

To understand crossmodal interactions, an effort must therefore be made in designing dual experimental tasks in order to force focus either on one or the other sense with concomitant contextual influence from the other sense (through endogenous attention; see Bertelson et al. [2000], Spence and Driver [1997].

Given that humans operate in a world of sensory uncertainty and that introspection fools us into thinking that perception is deterministic and certain cognitive biases may arise. These are undesirable in the experimental setting, because they may obscure the elucidation of other factors that influence the extraction of information about the world using biological sensors. These factors include ambiguity due to physical constraints (e.g. the mapping of 3D objects into 2D images, or the “aperture problem” in local motion detection [Castelo-Branco et al. 2000; 2002; 2007]), neural noise introduced in the early stages of sensory coding, and structural constraints on neural representations and computations [Knill and Pouget 2004]. It is therefore essential that psychophysical investigations of crossmodal interactions and perceptual decision take into account baseline uncertainty in contextual modulations. This has the important advantage of eliminating baseline perceptual bias.

Therefore, in the particular case of the interaction between vision and audition, we investigated crossmodal influences by requiring subjects to report only on motion perception conveyed either by audition (to test for visual contextual influences) or by vision (to test for auditory contextual influences), while receiving simultaneous input from both modalities. In this way the degree of perceptual uncertainty conveyed by the contextual stimuli and its importance on the final multisensory perception outcome is taken into account. Moreover, since it is commonly held that there is a tendency to attend preferentially toward vision, we specifically manipulated the visual contextual stimulus so as to maximise its ambiguity/unpredictability. This improved its comparability to the auditory contexts, in the sense that it would increase the detectability of any pattern of dominance of auditory context and also enable the isolation of contextual effects regardless of baseline perceptual bias.

Subjects were seated at about 36 cm viewing distance from a monitor flanked by two loudspeakers lying in the same plane as the monitor screen and at a distance of 56 cm from each other. Head position was stabilised using a chin/forehead rest. The experiment took place in a light- and sound-attenuated room in which the monitor and the loudspeakers were the only sources of light and sound, respectively. MATLAB and the purpose-built Psychophysical Software Toolbox Cogent 2000 were used for stimulus presentation and data collection.

As mentioned above, these experiments were designed either to force attentional focus on audition by requiring subjects to report only on motion direction of the audio stimulus while being presented a visual stimulus simultaneously, thus investigating the influence of the unattended (and ambiguous) visual stimulus on auditory motion perception, or vice-versa. Since the generalised belief is that spatially vision captures audition due to sensor reliability, visual stimuli were purposely designed to be ambiguous in terms of perceived motion coherence; more specifically, bistable grating (spatially coherent or incoherent) stimuli based on Alais et al. [1998] were used (see previous section and also below). This way, any visual or auditory capture phenomena would be clearly linked to crossmodal influences, and baseline perceptual bias factors could be ruled out.

The overall procedure (see Fig 3.3) was thus sequenced in three steps: (1) a pre-run with the purpose of establishing the most ambiguous visual condition (local grating angles yielding ambiguous coherent vs. incoherent percepts) for each subject and to quantify this ambiguity; (2) an experiment designed to test visual context induced biasing of auditory motion perception (global coherent motion, which emergence was unpredictable); (3) an experiment designed to test auditory (left or rightward motion) contextual influences on visual motion perception.

In any of the three steps, subjects had to report either answer A, answer B (the content of which — direction of auditory or visual motion — depended on the experiment, with A and B being mutually exclusive) or no answer when uncertain. Subjects were instructed to respond dynamically and interactively all throughout each trial, with key “Q” or key “P” kept pressed while answer A or answer B was valid, respectively, and no key pressed while subject would feel uncertain of what to respond. A timeline consisting of “A answered”, “B answered” and “no answer” periods would be saved on a log file, along with the percentages of total “A answered” time and “B answered” time relative to full trial length, per trial. A summary file consisting of a log of the

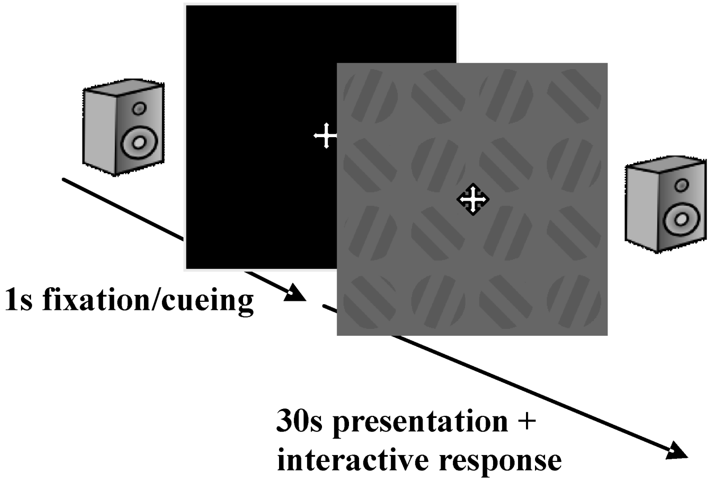


Figure 3.3: Illustration of sequence and timing of (visual and/or auditory) stimulus presentation for each trial. Local gratings may either be perceived as incoherent, or alternatively cohere into global leftward/rightward motion. Auditory motion was either leftward, rightward or absent.

total averages for each experiment of total "*A* answered" time and "*B* answered" time relative to full trial length would also be generated.

12 subjects were tested, 5 males and 7 females with ages comprehended between 20 and 59 years-old, with normal or corrected-to-normal visual acuity and normal auditory acuity. Informed consent was obtained following the guidelines of our local ethics committee.

In all experiments, stimuli were presented by a Dell Precision 380 computer on a 22-inch Mitsubushi monitor with 1024 × 768 resolution (see Fig. 1). The visual stimulus area size (95% of the screen) covered 24.7° visual angle and consisted of luminance-defined square wave gratings (grating period 1.1°) that could be manipulated independently in terms of orientation/direction; high luminance regions were measured as being 8.58 $^{\mathrm{cd}}/_{\mathrm{m}^2}$, while low luminance regions were measured as being 5.85 $^{\mathrm{cd}}/_{\mathrm{m}^2}$, hence Michelson contrast for these gratings was fixed to around 20%. Gratings were seen through 16 small circular apertures of 2.6° visual angle, arranged in a 4×4 array, with each aperture equally spaced from its vertical and horizontal neighbour. A fixation point was located at the centre of the display. The distance between aperture centres was 6.5° visual angle. Visual displays consisted of 31 frames displayed at a frame rate of 31.25 $^{\mathrm{frames}}/_{\mathrm{s}}$ for 30 s. The two independently moving gratings could be seen through the apertures with each grating occupying alternate diagonals of the array. Grating motions could be directed along 2 overall visual movement directions ("left", "right"). The grating directions/orientations might be symmetrical ($\pm 23^o, \pm 45^o$ and $\pm 68^o$ with respect to the vertical or horizontal axis) and asymmetrical ($-23^o/+45^o$ with respect to the vertical or horizontal axis), thus providing a total of 4 different orientation conditions. The gratings were always matched and differed only in orientation and direction.

As described by Alais et al. [1998], these stimuli resulted in bistable perception: a) when the gratings in the 16 apertures were reported as being seen to move coherently by the subject, a grouping process occurred, whereby they were perceived as four sets of complete (but partially occluded), concentric diamonds moving as if on a single surface behind the apertures; b) when incoherent motion was reported to be seen, the gratings appeared to move separately in orthogonal incoherent diagonal streams. Thus, an experimental pre-run was conducted, in order to determine optimal ambiguity levels prior to the contextual experiments — each trial lasted approximately 31 s, composed by 1 s of fixation/cueing and 30 s of stimulus presentation, throughout which the subjects were required to respond (see Fig. 3.3). Given the presentation of 4 orientation × 2 context types (8 trial types), total experimental time was about 4 mins. The pre-run experiment helped establish a high — $\approx 50\%$ — ambiguity of visual stimuli (see below). Motion of auditory stimuli were relatively less ambiguous and all included subjects could in fact easily determine is direction of motion in the absence of a context.

Subjects were required to report on visual motion, by pressing "Q" only when they perceived coherent motion and "P" only when they perceived more than one motion signal, and were asked to refrain from responding when unsure. This experiment enabled determination of single subject optimal ambiguity settings.

Then, the first contextual experiment consisted of investigating the crossmodal influence of a moving visual stimulus, based on the local grating orientations which elicited the most ambiguous perception from the subject (i.e. the most balanced coherent-incoherent motion response times, as determined in the pre-run experiment, which aimed to define an ambiguous context, with uncertain/unpredictable presence of coherent visual motion), on the perception of motion of a non-stationary auditory stimulus.

Visual stimuli were the same as in the prerun (defining optimal ambiguity[1] settings), but this time using only the chosen grating orientation consistent with the two possible coherent global motion horizontal directions (left/right), resulting in 2 differ-

[1] i.e. local aperture grating orientation.

ent visual conditions plus a “no visual stimuli” control condition. Auditory stimuli were presented at around 60 dB SPL (sound pressure level) at listening distance. They consisted of 30 s of mono broadband noise processed in stereo. Sound was generated so as to simulate two different motions, left-to-right or right-to-left, so that the auditory motion signal direction could be consistent or inconsistent with the visually defined motion signal in any particular instant - two mirrored audio conditions were created, with the simulated sound-source alternating between 5 s-duration left-to-right and right-to-left motions so that each direction would encompass 50% of trial run-time. Auditory and visual motion speeds were designed to be approximately coincident, moving at 2^{o} visual angle per second.

Sound-source motion was simulated using an amplitude panning technique, adjusting the amplitude of the signal being delivered to each loudspeaker to simulate the directional properties of interaural level difference (ILD) cues (refer to section 2.2.2 for complete definition). The sound files were presented binaurally using the tangent law introduced by Bennett [Pulkki 1998]

$$\frac{\tan \phi}{\tan \theta_0} = \frac{g_l - g_r}{g_l + g_r} \tag{3.1}$$

with g_l and g_r being the gain factors scaling the amplitude of the signal applied to the right and left loudspeaker, respectively. The equation can be manipulated by assuming a constant virtual volume level $C = 1$. This can be accomplished by ensuring that $g_l + g_r = 1$, since keeping the virtual source volume level constant approximates a constant distance from the sound-source.

Two runs of 30 trials, separated by a resting period to avoid subject fatigue, were run for each visual-auditory stimulus combination (visual motion direction * auditory motion direction), resulting in $5 + 5$ random sequences of $2 * 3$ conditions $= 6$ trials of 30 s + 1 s each (with the same presentation sequence and timing of the prerun, shown on Fig. 3.3), resulting in approximately 31 min total running time of the experiment. Subjects were required to report on auditory motion, by pressing “Q” only when they perceived right-to-left motion and “P” only when they perceived left-to-right motion, and were asked to refrain from responding when unsure.

The second experiment consisted of investigating the crossmodal influence of a moving auditory stimulus on the perception of motion of a non-stationary ambiguous visual stimulus, based on the grating local orientation which elicited the most ambiguous perception from the subject (i.e. the most balanced coherent-incoherent motion perception) as the outcome of the prerun.

Visual stimuli were the same as in experiment 1; auditory stimuli were created using the same techniques described above; for this experiment, however, the 2 auditory conditions were designed using only one audio motion direction per condition. Since the presentation time is 30 s-long, and a complete (back and forth) motion path takes 10s, the audio streams were designed so as to restart the sound-source motion path twice. An additional "no audio" control condition was also designed.

Again, two runs of 30 trials, separated by a resting period to avoid subject fatigue, were run for each visual-auditory stimulus combination (visual motion direction * auditory motion direction), resulting in 5 + 5 random sequences of 3 conditions * 2 conditions = 6 trials of 30 s + 1 s each (with the same presentation sequence and timing of the prerun and experiment 1, shown on Fig. 3.3), resulting in approximately 31 min total running time of the experiment. Subjects were required to report on visual motion presence or absence of global directional coherence as on the prerun.

Statistical analysis of experimental data was performed resorting to repeated measures General Linear model analyses, after verifying normality and homogeneity of measures, using SPSS 16. It was found that global visual context did significantly modulate perceived direction of auditory signals in spite of their relatively low ambiguity. Repeated measures GLM analysis did indeed reveal a main effect of visual context ($p << 0.0001$).

Post hoc analyses of the sources of these effects showed that these effects were significant both for rightward coherence in visual contexts ($p < 0.001$) and leftward coherence ($p < 0.002$). Interestingly, this *post hoc* analysis also revealed that the sources of these effects were mainly due to congruent *vs* incongruent visual contexts ($p = 0.003$ for leftward visual coherence and $p = 0.001$ for rightward visual coherence contextual conditions). Fig. 3.4 summarises the main effects of visual context on auditory perception.

To prove that the effect of visual context was due to global motion perception and not significantly influenced by the local orientation features of the stimuli we run a control analysis to probe whether local orientation changed perceived auditory direction. No significant effects of local visual properties on mean perceived auditory direction were found Fig. 3.5.

Surprisingly, we have found that auditory context did not significantly modulate perceive coherence of spatially distributed visual signals ($p = 0.5, ns$), suggesting that even when auditory signals have relatively lower ambiguity they may not be sufficient to modulate visual motion integration of local/global bistable stimuli (Fig. 3.6). This was further confirmed by *post hoc* analysis that did not reveal any auditory contextual effect on perceived coherent visual global motion.

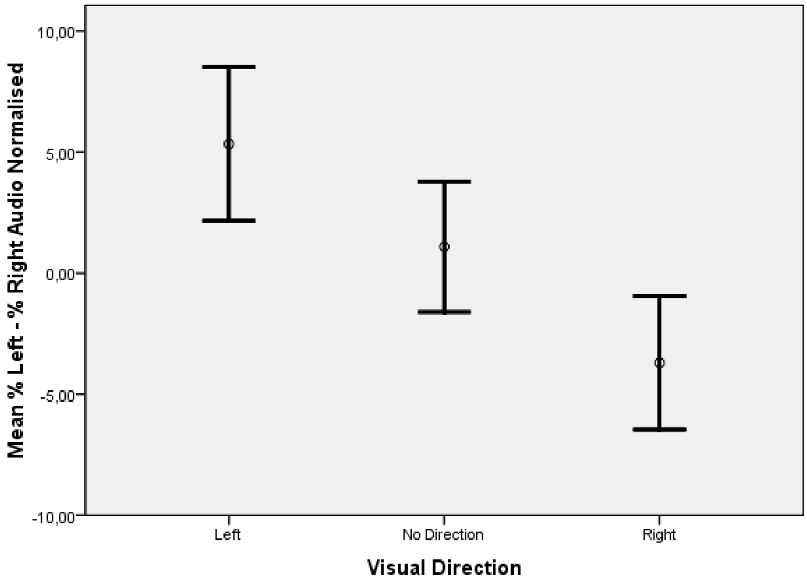


Figure 3.4: Effect of visual context on auditory motion perception. When attention is focused on auditory attributes and visual stimuli play a significant contextual modulation is observed. Ordinate: % Left - % Right Audio Normalised = (% Left Audio - % Right Audio)/(% Left Audio + % Right Audio). Abscissa: contextual visual motion conditions.

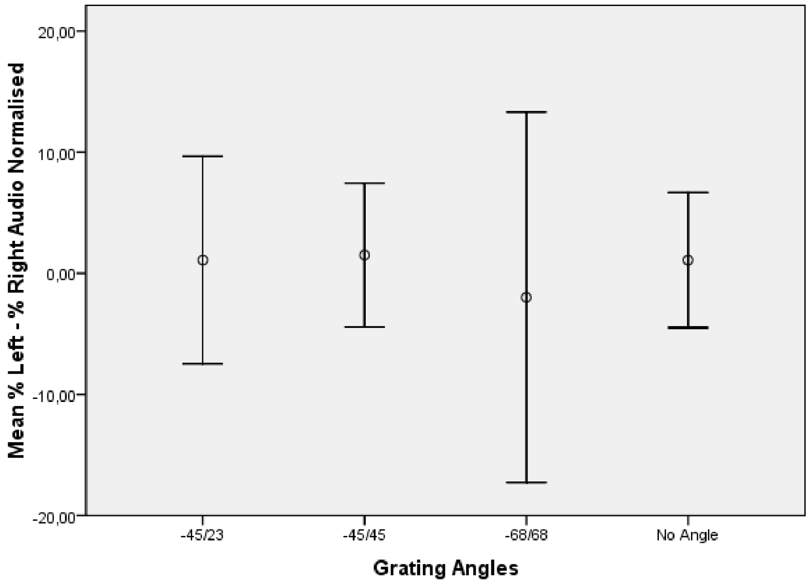


Figure 3.5: Effect of local orientation differences in visual stimuli on perceived auditory motion — local orientation differences do not significantly change perceived auditory motion. Ordinate measures calculated as in Fig. 3.4.

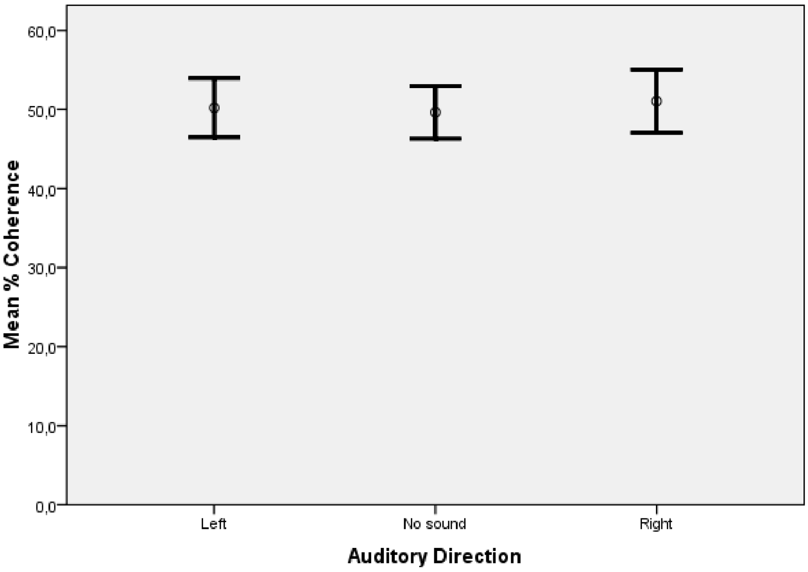


Figure 3.6: When attention is focussed on visual attributes and auditory stimuli play a contextual role no significant modulation is observed, in spite of the lower ambiguity of auditory stimuli as compared to visual stimuli (note the high — 50% — ambiguity of visual stimuli).

3.3 Conclusions

In the first baseline study, concerning visual motion, we have shown that humans show processing bias for grouping of ambiguous moving stimuli in the horizontal direction as opposed to the vertical direction. Right-left asymmetries were specifically present only for discontinuous cues. Our results do thereby suggest an important Bayesian prior in human perception, namely that horizontally moving cues are more likely to belong to a single coherent object, in particular a potential predator.

Subjects perceive spatially distributed stimuli moving from left to right as more coherent than vice versa, indicating a functional asymmetry that is presumably due to our frequent exposure to spatially distributed letter strings in reading. The most remarkable finding of this study was however that significantly more perceived coherence is observed for horizontally moving stimuli (right and left) compared to vertically moving stimuli (up and down). This indicates that humans divide the world into axes of symmetry in which stimuli along the horizontal plane are grouped more often than stimuli in the vertical plane. These findings extend the notion of “psychophysical performance fields” proposed by Carrasco et al. [2001], to a novel concept of “perceptual decision field”.

We conclude that a simple Bayesian-like rule for perceptual decision is at work, based on the fact that most animate objects in human philo- and ontogenetic experience move along the horizon: horizontally moving cues are more likely to belong to a single coherent object, namely a potential predator.

In the second baseline study, concerning visual- and auditory-based motion perception given auditory and visual context, respectively, we have shown that even when visual motion is unpredictably bistable (see also Alais et al. [1998]) it still contextually dominates auditory motion perception in an asymmetric manner. This evidence for asymmetric dominance of visual *vs* auditory context in perceptual decision is relevant to understand a current discussion on the neural weights across modalities that determine perceptual decision [Soto-Faraco et al. 2003, Knill and Pouget 2004, Burr and Alais 2006].

Although one cannot exclude that symmetric weights may still be present at the sensory level, our paradigm, whereby sensory stimulation was identical but the role of the context varied according to task instruction an attentional focus on each modality, clearly shows that the visual modality dominates when the higher level perceptual decision domain is involved The fact that our subjects could not predict *a priori* when moments of motion coherence would emerge provides evidence that baseline decision biases are not explaining the observed visual contextual dominance on auditory motion perception.

It is quite surprising that in spite of the fact that auditory motion was relatively less ambiguous and more predictable to our subjects, it still yielded non significant modulation under these physically matched conditions of visual and auditory motion. Absent effects were documented even under congruent conditions, in stark contrast with visual modulatory effects, which were particularly strong when congruent *vs* incongruent conditions were contrasted.

In conclusion, perceptual decision under contextual modulation is dominated by the visual over the auditory modality even when visual global motion signals are ambiguous or unpredictable. Future studies on the neural mechanisms of perceptual decision should address the implications of these findings for high level Bayesian models, such as the one proposed in chapter 2, of crossmodal interactions in motion perception, by applying the experimental paradigm presented on Fig. 2.1.

Chapter 4

Implementation of Active Exploration Using Bayesian Models for Multimodal Perception

4.1 Active Exploration Using the Bayesian Volumetric Map

The availability of a probabilistic framework to implement spatial mapping of the environment substantiated by the BVM allows the use of the concept of *information entropy*, which can be used to promote an exploratory behaviour of areas of the environment corresponding to cells on the volumetric map associated to high uncertainty, an idea explored by Rocha, Dias, and Carvalho [2005a;b].

Information in the BVM is stored as the *probability of each cell being in a certain state*, defined in the BP of Fig. 2.10 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:

$$H(c) \equiv H(V_c, O_c) = -\sum_{\substack{o_c \in \mathcal{O} \\ v_c \in \mathcal{V}}} P(v_c\, o_c|z\, c) \log P(v_c\, o_c|z\, c) \qquad (4.1)$$

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

$$\vec{\nabla} H(c) \approx [H(c) - H(c_{\rho -}), H(c) - H(c_{\theta -}), H(c) - H(c_{\phi -})]^T \tag{4.2}$$

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

A great advantage of the BVM over implementations of occupancy maps using regular partitioning of Euclidean space is the fact that the log-spherical configuration avoids the need for time-consuming ray-casting techniques when computing a gaze direction for active exploration, since the log-spherical space is already defined based on directions (θ, ϕ). Hence, the active exploration algorithm is simplified to the completion of the following steps (see Fig. 4.1):

1. Find the last non-occluded, close-to-empty (i.e. $P([O_C = 1]|[C = c]) < .5$) cell for the whole span of directions $(\theta_{\max}, \phi_{\max})$ in the BVM — these are considered to be the so-called *frontier cells* as defined by Rocha et al. [2005a]; the set of all frontier cells will be denoted here as $\mathcal{F} \in \mathcal{Y}$.

2. Compute the joint entropy gradient for each of the frontier cells and select $c_s = \arg\max_{c \in \mathcal{F}} \left[(1 - P([O_C = 1]|[C = c])) \|\vec{\nabla} H(c)\| \right]$ as the best candidate cell to direct gaze to. In case there is more than one global maximum, choose the cell corresponding to the direction closest to the current heading, so as to ensure minimum gaze shift rotation effort.

3. Compute gaze direction as being (θ_C, ϕ_C), where θ_C and ϕ_C are the angles that bisect cell $[C = c_s]$ (i.e. which pass through the geometric centre of cell c_s).

The full BVM entropy-based active perception system is described by the block diagram presented in Fig. 4.2.

The BVM is extendible in such a way that other properties, characterised by additional random variables and corresponding probabilities might be represented, other than the already implemented occupancy and local motion properties O_C and V_C, by augmenting the hierarchy of operators through Bayesian subprogramming — see Appendix A on page 121.

Therefore, we introduce a new random variable U_C, which takes the algorithm presented above and expresses it in a compact mathematical form:

$$U_C = \begin{cases} (1 - P([O_C = 1]|C)) \frac{\|\vec{\nabla} H(C)\|}{\max \|\vec{\nabla} H(C)\|} & C \in \mathcal{F}, \\ 0 & C \notin \mathcal{F}. \end{cases} \tag{4.3}$$

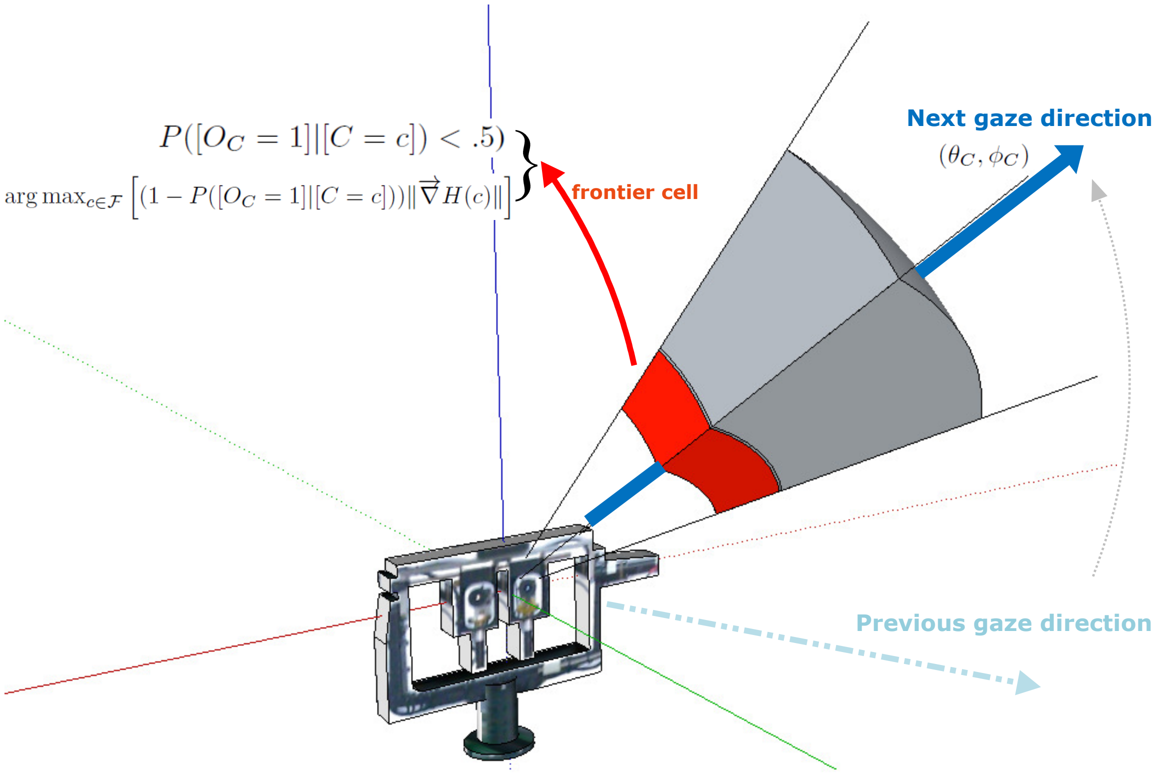

$P([O_C = 1] | [C = c]) < .5)$
$\arg\max_{c \in \mathcal{F}} \left[(1 - P([O_C = 1] | [C = c])) \| \vec{\nabla} H(c) \| \right]$
frontier cell
Next gaze direction
(θ_C, ϕ_C)
Previous gaze direction

Figure 4.1: Illustration of the entropy-based active exploration process using the Bayesian Volumetric Map. The result of applying the algorithm steps described in the main text is depicted. When there exists more than one maximum for $(1 - P([O_C = 1] | [C = c])) \| \vec{\nabla} H(c) \|$, the frontier cell corresponding to the direction closest to the current heading is chosen, so as to ensure minimum gaze shift rotation effort.

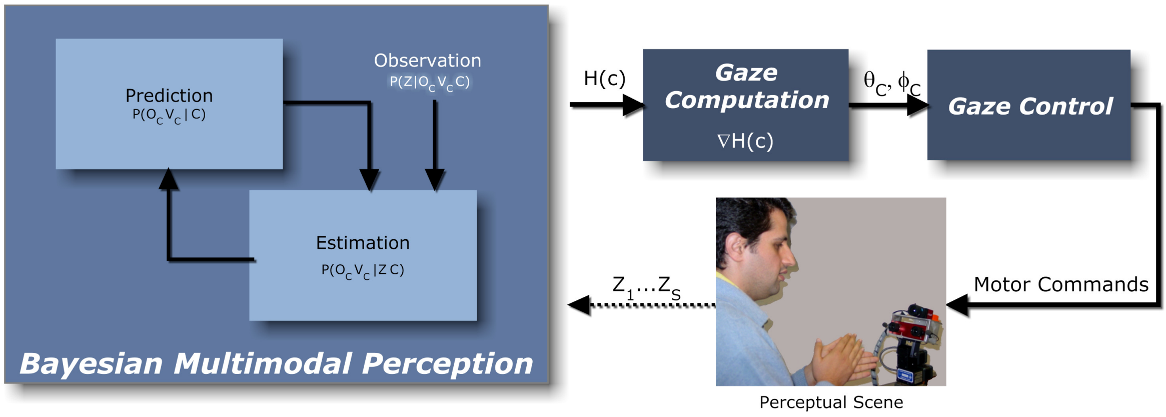


Figure 4.2: Active multimodal perception using entropy-based exploration. The "Gaze Control" module has been described elsewhere — see list of publications in Appendix C — and is beyond the scope of this text.

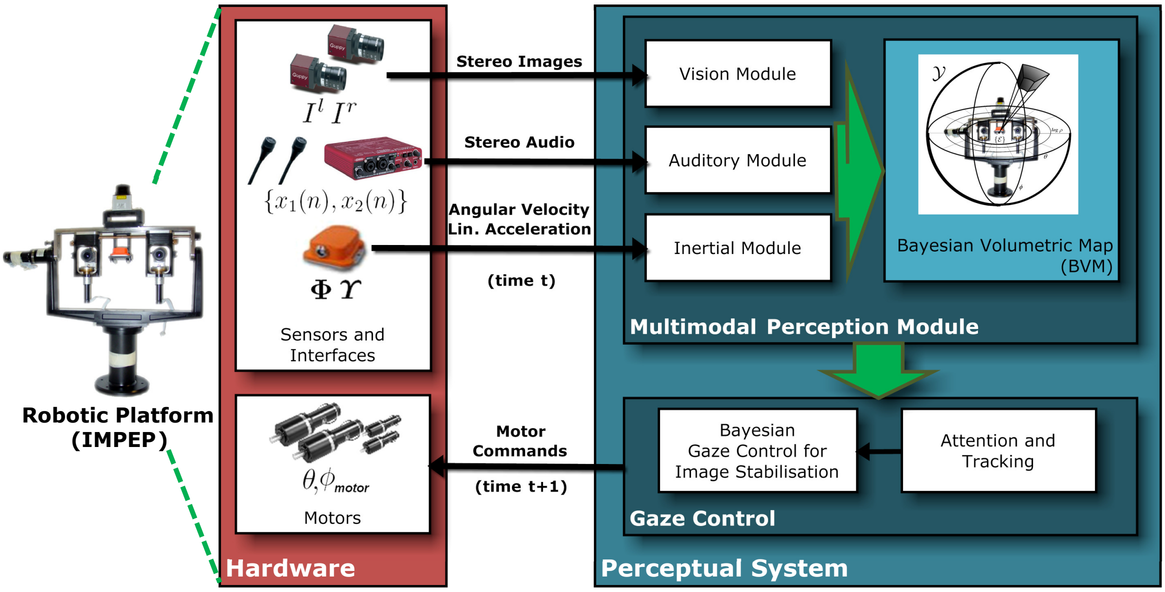


Figure 4.3: Integration layout for the active multimodal perception system.

4.2 System Implementation and Calibration

4.2.1 Overall implementation details

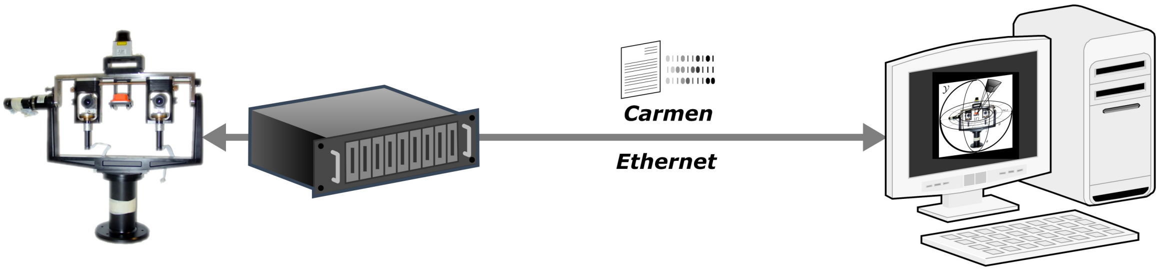
The real-time active multimodal perception system integrates the Bayesian framework described in chapter 2 and computes a stabilised gaze shift towards a site on the environment to be explored in the subsequent time step — Fig. 4.3 shows an overview of the system's layout and integration.

The BVM-IMPEP real-time system was developed using the following software:

- **Vision sensor system**: With the OpenCV toolbox and the implementation by Gallup [2009] of a basic binocular stereo algorithm on GPU using CUDA. The algorithm reportedly runs at 40 Hz on 640 × 480 images at 50 disparities, computing left and right disparity maps and performing left-right consistency validation (which in our adaptation is used to produce the stereovision confidence maps).

- **Binaural sensor system**: Using an adaptation of the real-time software kindly made available by the Speech and Hearing Group at the University of Shefield [Lu et al. 2007] to implement binaural cue analysis as described on chapter 2.

- **Bayesian Volumetric Map, Bayesian sensor models and active exploration**: using our proprietary, parallel processing, single-precision GPU imple-



- Mobile DualCore Intel Pentium M, 1666MHz
- Cache L1 (32KB) and L2 (2048KB)
- 1024 MB RAM
- 74GB hard-disk
- Integrated graphics card

- DualCore Intel Pentium D 950, 3.40GHz
- Cache L1 (32KB) and L2 (2048KB)
- 1 GB RAM
- 80GB, 7200 rpm hard-disk
- PCI-Express NVIDIA GeForce 9800 GTX (512MB)

Figure 4.4: BVM-IMPEP system network diagram.

mentation developed with NVIDIA's CUDA, described on section 4.2.2.

The BVM-IMPEP system is composed of a local Ethernet network comprised of two PCs communicating and synchronising via Carmen messaging [Montemerlo et al. 2007], one for all the sensory and BVM framework processing (including CUDA processing on a NVIDIA GeForce 9800 GTX, compute capability 1.1), and the other for controlling the IMPEP head motors, designed for portability (i.e. low-consumption and lightweight) in order to be mounted on mobile robotic platforms in the future — see Fig. 4.4. Both are equipped with Ubuntu Linux v9.04.

4.2.2 Bayesian Volumetric Map implementation on GPU using CUDA

The activity diagram for the BVM Bayesian framework is presented on Fig. 4.5, depicting an inference step corresponding to time t and respective timeline. In the following lines, our GPU implementation of the BVM algorithms developed with NVIDIA's CUDA that exectute this timeline will be described in more detail.

Bayesian Volumetric Map filter

The BVM filter, which comprises the processing lane on the right of Fig. 4.5, launches kernels based on a single three-dimensional grid corresponding to the log-spherical configuration — see Fig. 4.6. In fact, both input matrices (i.e. observations and previous system state matrices) and output matrices (i.e. current state matrices) have the same indexing system. Blocks on this grid were arranged in such a way that their 2D indices would coincide with azimuth θ and elevation ϕ indices on the grid,

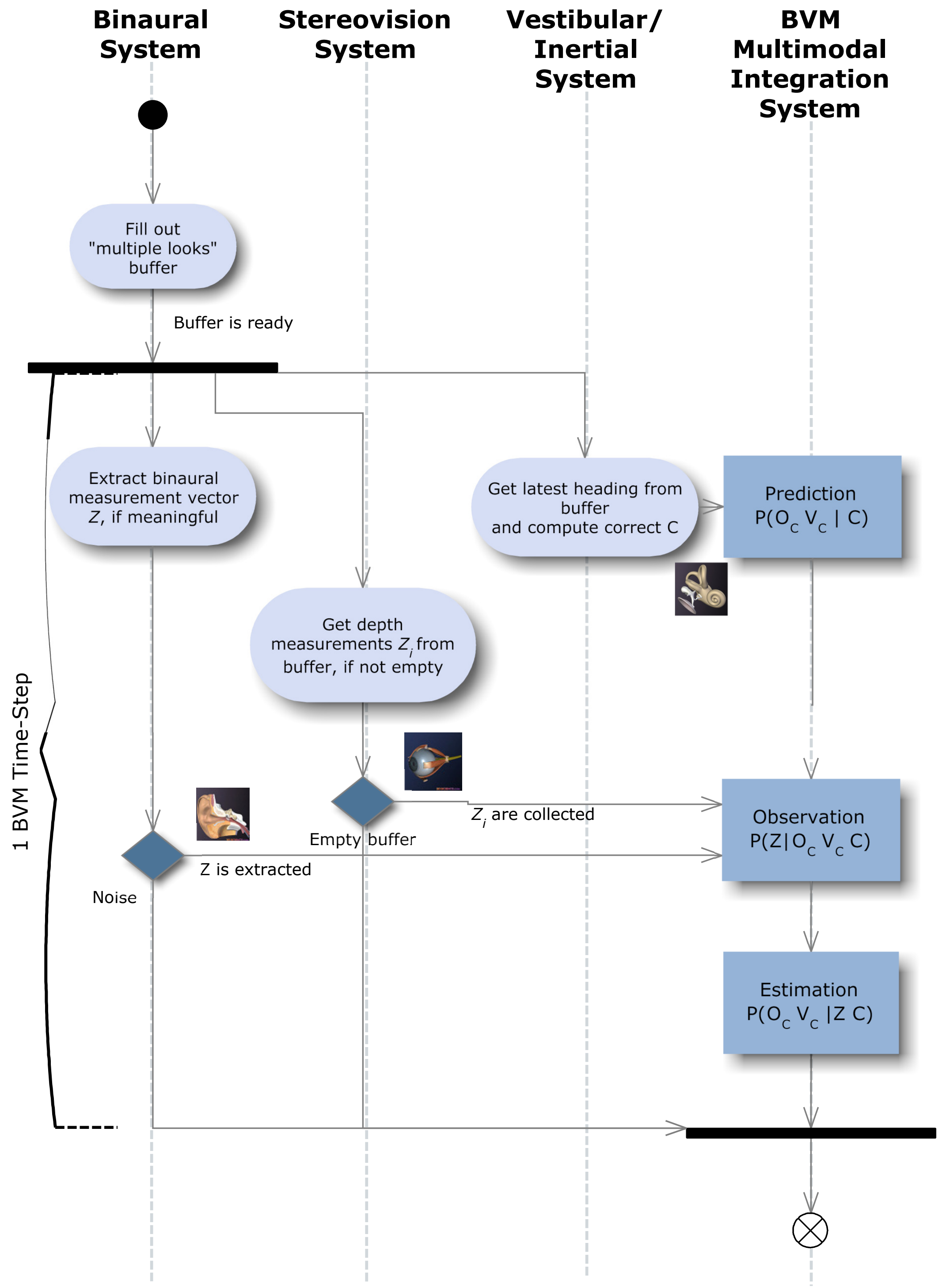

Binaural System
Stereovision System
Vestibular/ Inertial System
BVM Multimodal Integration System
Fill out "multiple looks" buffer
Buffer is ready
Extract binaural measurement vector Z, if meaningful
Get latest heading from buffer and compute correct C
Prediction P(O_C V_C | C)
Get depth measurements Z_i from buffer, if not empty
Z_i are collected
Empty buffer
Observation P(Z|O_C V_C C)
Z is extracted
Noise
Estimation P(O_C V_C |Z C)
1 BVM Time-Step

Figure 4.5: Activity diagram for an inference time-step at time t. Each vertical lane represents a processing thread of the module labelled in the corresponding title. Maximum processing times (for $N = 10, \Delta\theta = 1^o, \Delta\phi = 2^o$) are also presented in the timeline for reference.

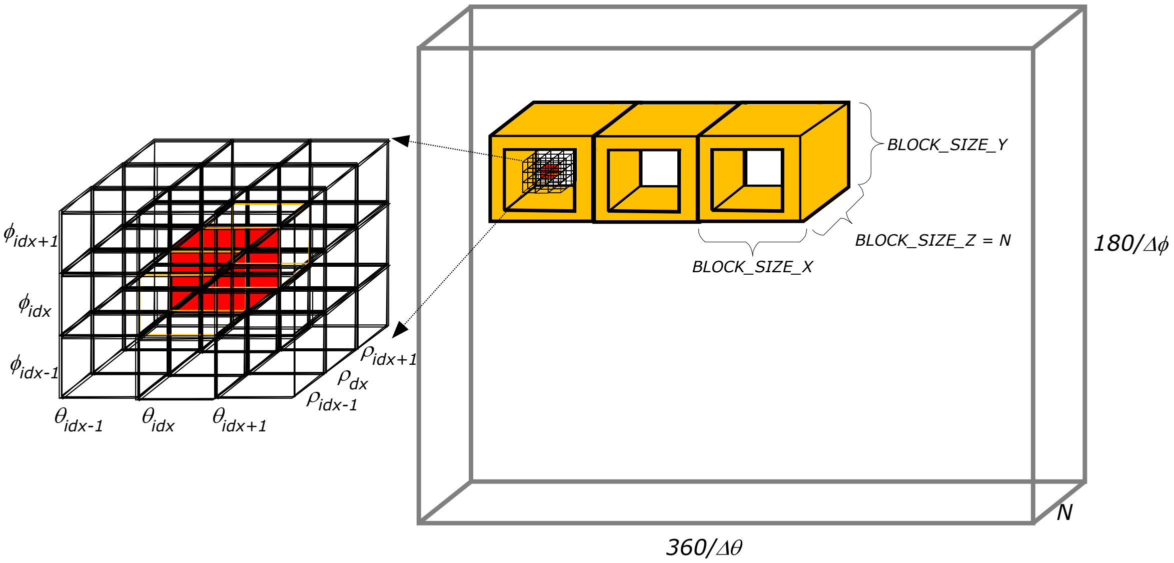


Figure 4.6: BVM filter CUDA implementation. On the right, the overall 3D BVM grid is shown. On the left, a zoom in on the 9 adjacent cells needed to update a central cell of the BVM are shown — this means that shared memory is required. As mentioned before, CUDA allows reference to each thread using a three-dimensional index; however, it only allows two-dimensional indexing for thread blocks. For this reason, we decided to assign the smallest dimension to the third axis (from now on referred to as "depth" — with size N), and by making all blocks the same depth as the global grid — this ensures that the block two-dimensional index corresponds to the remaining axes, simplifying memory indexing computations. Each thread loads its cell's previous state into shared memory and the log-probabilities for sensor measurements. The need for access to the previous states of adjacent cells further complicates the implementation by forcing the use of *aprons*, depicted in yellow within the thread blocks (see Fig. 4.8(a) for further details on kernel implementation using aprons).

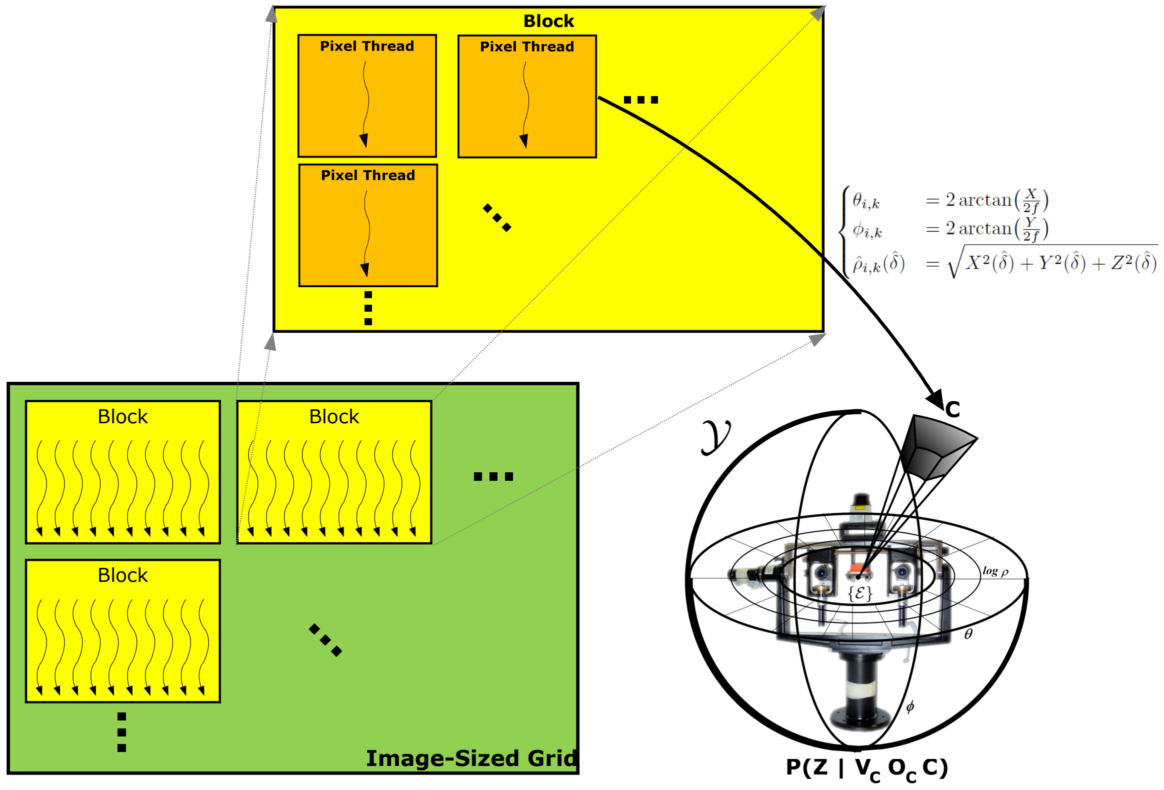
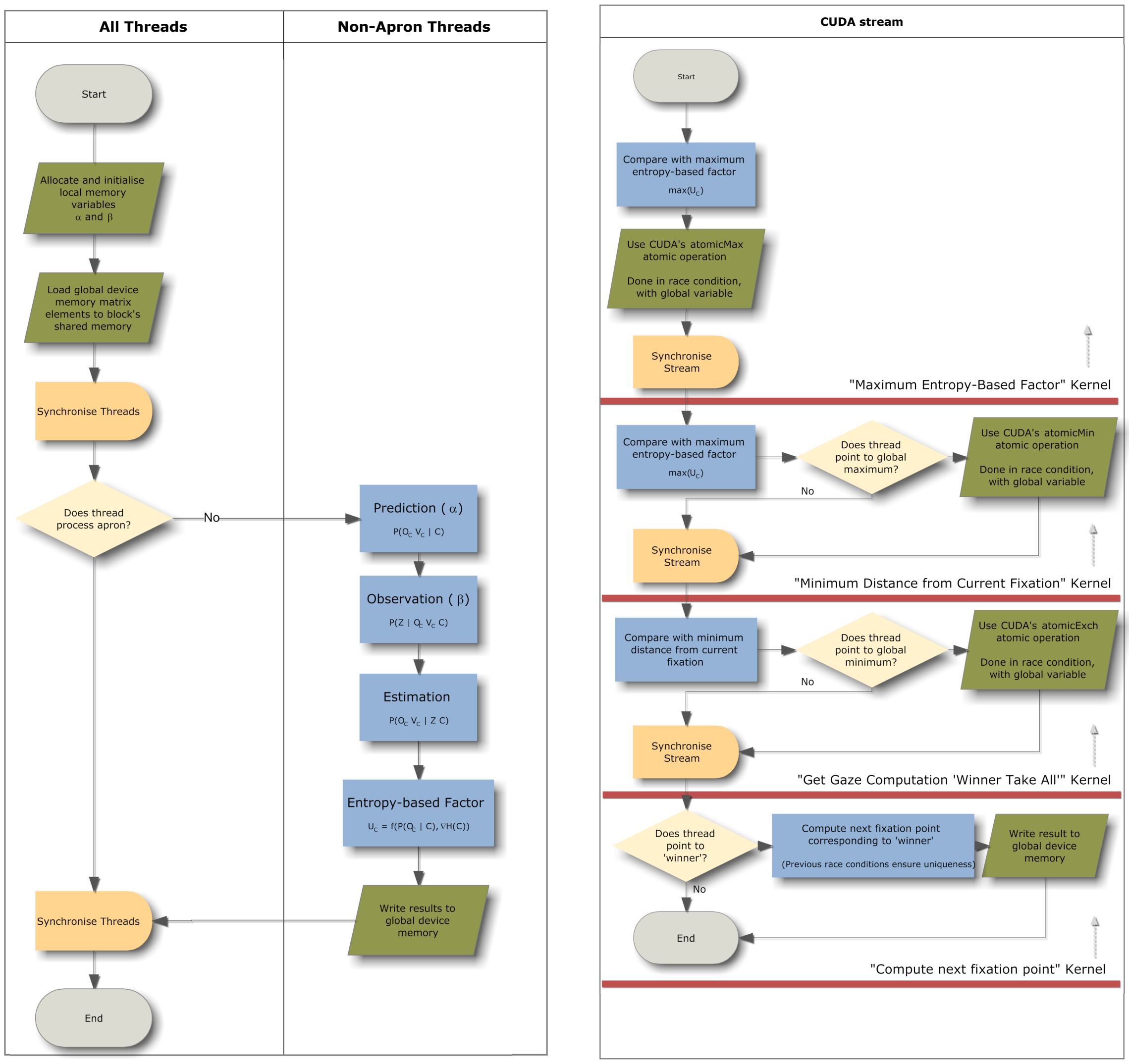

Block
Pixel Thread
Pixel Thread
Pixel Thread
$\theta_{i,k} = 2\arctan(\frac{X}{2f})$
$\phi_{i,k} = 2\arctan(\frac{Y}{2f})$
$\hat{\rho}_{i,k}(\hat{\delta}) = \sqrt{X^2(\hat{\delta}) + Y^2(\hat{\delta}) + Z^2(\hat{\delta})}$
Block
Block
Block
Image-Sized Grid
C
$\mathcal{Y}$
log ρ
{ℰ}
θ
ϕ
P(Z | V_C O_C C)

Figure 4.7: Stereovision sensor model CUDA implementation. Each thread independently processes one pixel of the egocentric-referred depth map and confidence images (no use of shared memory required), computes the corresponding cell C on the BVM log-spherical spatial configuration using the equation shown, and updates two data structures in global device memory with that configuration storing log-probabilities corresponding to $P(Z|[O_C = 1]\, C)$ and $P(Z|[O_C = 0]\, C)$ (independent of velocity V_C), respectively. The update is performed using atomic summation operations provided by CUDA compute capability 1.1 and higher [NVIDIA 2007]. Atomic operations are needed due to the many-to-one correspondence between pixels and cells on the BVM; however, the order of summation is, obviously, nonimportant. Finally, since all atomic operations except "exchange" only accept integers as arguments, log-probabilities are converted from to floating-point to integer through a truncated multiplication by 10^n, with n corresponding to the desired precision (in our implementation, we used $n = 4$).


All Threads
Non-Apron Threads
Start
Allocate and initialise local memory variables α and β
Load global device memory matrix elements to block's shared memory
Synchronise Threads
Does thread process apron?
No
Prediction (α)
P(O_C V_C | C)
Observation (β)
P(Z | O_C V_C C)
Estimation
P(O_C V_C | Z C)
Entropy-based Factor
U_C = f(P(O_C | C), ∇H(C))
Write results to global device memory
Synchronise Threads
End
CUDA stream
Start
Compare with maximum entropy-based factor
max(U_C)
Use CUDA's atomicMax atomic operation
Done in race condition, with global variable
Synchronise Stream
"Maximum Entropy-Based Factor" Kernel
Compare with maximum entropy-based factor
max(U_C)
Does thread point to global maximum?
Use CUDA's atomicMin atomic operation
Done in race condition, with global variable
No
Synchronise Stream
"Minimum Distance from Current Fixation" Kernel
Compare with minimum distance from current fixation
Does thread point to global minimum?
Use CUDA's atomicExch atomic operation
Done in race condition, with global variable
No
Synchronise Stream
"Get Gaze Computation 'Winner Take All'" Kernel
Does thread point to 'winner'?
Compute next fixation point corresponding to 'winner'
(Previous race conditions ensure uniqueness)
Write result to global device memory
No
End
"Compute next fixation point" Kernel

(a) BVM filter CUDA kernel flowchart. *Aprons* are the limiting cells of the block, to which correspond threads that cannot access adjacent states, and therefore with the sole mission of loading their respective states into shared memory — thus, blocks must overlap as their indices change, so that all cells have the chance to be non-apron. After all threads, apron or non-apron, load their respective previous states into shared memory, all non-apron threads then perform Bayesian filter estimation and update the states, as depicted. The "Observation" box here denotes the computation of β by multiplying all available outputs from the stereovision and binaural Bayesian sensor models denoted as the "Observation" box of Fig. 4.5.

(b) Active exploration CUDA stream flowchart. Four consecutive kernels in a sequential CUDA stream were used to implement the active exploration algorithm. The division of the processing workload into separate kernels was necessary due to the fact that the only way to enforce synchronisation between all concurrent CUDA threads in a **grid** (as opposed to all threads in a **block**, which is only a subset of the former) is to wait for all kernels running on that grid to exit — this is only possible at CUDA stream level (see main text for the definition of CUDA stream). CUDA atomic operations (refer to Fig. 4.7 for more information) and global memory were used to pass on data from one kernel to the next without the need for additional memory operations.

Figure 4.8: BVM CUDA implementation flowcharts.

assuming that the full N-depth of the log-distance index is always copied to shared memory.

By trial-and-error we arrived at the conclusion that block size was limited by shared memory resources to $5 \times 5 \times N$ for $N \leq 10$ and $3 \times 3 \times N$ for $N = 11$, which would therefore be the top limit for depth using this rationale of a single grid for the whole BVM space. In fact, for $N < 11$, there were 250 threads and 8000 bytes of shared memory per block, thus limiting the maximum number of blocks per multiprocessor to 2 for the compute capability 1.1 of the GeForce 9800 GTX; for $N = 11$, on the other hand, there were 90 threads and 2880 bytes of shared memory per block, increasing the limit of blocks per multiprocessor to 5.

The flowchart for the BVM filter kernel is shown on Fig. 4.8(a).

Bayesian processing of stereovision and binaural data

The stererovision sensor model, which comprises the second processing lane from the left and the "Observation" box of Fig. 4.5, launches kernels based on two-dimensional grids corresponding to image configuration — see Fig. 4.7. In fact, its input matrices (left and right images, and disparity and confidence maps) have the same indexing system, while its output matrices (visual observation matrices) have the same indexing as the BVM grid of Fig. 4.6.

By trial-and-error we arrived at the conclusion that block size was limited by register memory resources to 16×16 for 640×480 images. This also ensured that it was a multiple of the warp size so as to achieve maximum efficiency.

The implementation of the binaural sensor model, corresponding to the processing lane on the left and the "Observation" box of Fig. 4.5, contrastingly, is very simple — a vector of binaural readings is used as an input and a grid as shown on Fig. 4.6, but without resorting to aprons (i.e. shared memory; see Figs. 4.6 and 4.8(a) for a detailed explanation of this notion), was used to update sensor model measurement data structures analogous to those of the stereovision sensor model, by referring to a lookup table with normal distribution parameters taken from the auditory system calibration procedure (see section 4.2.3).

When there are visual and binaural measurements available simultaneously, two CUDA streams[1] are created (i.e. forked), one for each sensor model, and then destroyed (i.e. merged).

[1] CUDA *streams* are concurrent lanes of execution that allow parallel execution of multiple kernels on the GPU.

Active exploration

The active exploration algorithm was implemented resorting to CUDA atomic operations, global memory and four consecutive kernels in a sequential CUDA stream. This implementation is detailed on Fig. 4.8(b).

To avoid the adverse effects of motion blur on stereoscopic measurements, a strategy similar to what is adopted by the human brain is implemented for fixations and gaze shifts — fixation is accomplished by processing data coming from the stereovision system for a few hundred milliseconds [Carpenter 2004; 2000, Caspi et al. 2004], followed by a process similar to the so-called *saccadic suppresion*, in which the magnocellular visual pathway (mainly supplying data to the dorsal pathway, which we intend to model) is actively suppressed during saccades [Burr et al. 1994, Watson and Krekelberg 2009]. In our parallel of this process, we simply halt gaze shift generation for a few iterations of the vision sensor model updates (thus simulating fixation), and then stop the updates during gaze shifts (thus simulating saccadic suppresion), without stopping the low-level processing of the stereovision system, which might be used in the future for other purposes.

4.2.3 System calibration

Visuoinertial calibration

Accurate camera calibration can greatly simplify solutions to many important vision problems such as the stereo vision problem, the three-dimensional visual tracking problem, the mobile-robot visual guidance problem, the 3D reconstruction problem, the 3D visual information registration problem, etc. For example, it is well known that a well-calibrated stereo vision system would not only dramatically reduce the complexity of the stereo correspondence problem but also significantly reduce the 3D estimation error [Shih, Hung, and Lin 1998].

Camera calibration can be performed using a standard stereovision calibration software to estimate left and right camera *intrisic parameters* (i.e. focal length and distortion parameters for undistorting images for processing) and *extrinsic parameters* (i.e. transformation between camera local coordinate systems — in the case of an ideal frontoparallel setup, the estimation of baseline b) that allow the application of the reprojection equation:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & f \\ 0 & 0 & \frac{1}{b} & 0 \end{bmatrix} \begin{bmatrix} u_l - \frac{\hat{\delta}}{2} \\ v_l \\ \hat{\delta} \\ 1 \end{bmatrix} = \begin{bmatrix} WX \\ WY \\ WZ \\ W \end{bmatrix} \tag{4.4}$$

where u_l is the horizontal coordinate and v_l is the vertical coordinate of a point on the left camera, 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} of equation (2.19) is defined as being equal to the maximum error exhibited by the stereovision system.

Finally, to determine $(\theta_{i,k}, \phi_{i,k})$ and $\hat{\rho}_{i,k}(\hat{\delta})$ (i.e. to perform the cartesian-to-spherical transformation) for each projection line (i,k) to use with the vision sensor model given in Figure 2.11, the following relations are built from equation (4.4),

$$\begin{cases} \theta_{i,k} & = 2\arctan\left(\frac{X}{2f}\right) \\ \phi_{i,k} & = 2\arctan\left(\frac{Y}{2f}\right) \\ \hat{\rho}_{i,k}(\hat{\delta}) & = \sqrt{X^2(\hat{\delta}) + Y^2(\hat{\delta}) + Z^2(\hat{\delta})} \end{cases} \tag{4.5}$$

Given $\theta_{i,k}$ and $\phi_{i,k}$, it becomes possible at any moment to compute depth from a given disparity estimate by substitution of the two first expressions onto the last in Equation 4.5, yielding

$$\hat{\rho}_{i,k}(\hat{\delta}) = f\sqrt{4\left(\tan^2\frac{\theta_{i,k}}{2} + \tan^2\frac{\phi_{i,k}}{2}\right) + \left(\frac{b}{\hat{\delta}}\right)^2} \tag{4.6}$$

In order to determine the rigid rotation between the INS frame of reference $\{\mathcal{I}\}$ and the right camera frame of reference $\{\mathcal{C}_R\}$, both sensors are used to measure the vertical direction[2]. When the IMU sensed acceleration is equal in magnitude to gravity, the sensed direction is the vertical. For the camera, using a specific calibration target such as a chessboard target placed vertically, the vertical direction can be taken from the corresponding vanishing point [Lobo and Dias 2007].

[2] The right camera was arbitrarily chosen as the dominant eye throughout this work — knowing the geometry of the stereovision system through calibration, relating $\{\mathcal{C}_R\}$ to $\{\mathcal{E}\}$ is trivial.

Rotation Calibration Summary

- N static observations at distinct positions:
  - vertical chessboard target;
  - save image and corresponding inertial data.
- Perform standard camera calibration for image set.
- Compute rotation quaternion:
  - use target vertical vanishing points ${}^{\mathcal{C}}\boldsymbol{v}_i$ detected for camera calibration;
  - inertial data from accelerometers provide ${}^{\mathcal{I}}\boldsymbol{v}_i$;
  - rotation quaternion $\mathring{\mathbf{q}}$ given by (4.8).

Figure 4.9: Summary of required steps to perform calibration of rotation between camera and IMU using the proposed algorithm.

Let ${}^{\mathcal{I}}\boldsymbol{v}_i$ be a measurement of the vertical by the inertial sensors, and ${}^{\mathcal{C}_R}\boldsymbol{v}_i$ the corresponding measurement made by the camera derived from some scene vanishing point. We want to determine the unit quaternion $\mathring{\mathbf{q}}$ that rotates inertial measurements in the inertial sensor frame of reference $\{\mathcal{I}\}$ to the camera frame of reference $\{\mathcal{C}_R\}$; in other words, we want to find the unit quaternion $\mathring{\mathbf{q}}$ that maximises [Lobo and Dias 2007]

$$\sum_{i=1}^{n} (\mathring{\mathbf{q}}\, {}^{\mathcal{I}}\boldsymbol{v}_i\, \mathring{\mathbf{q}}^*) \cdot {}^{\mathcal{C}_R}\boldsymbol{v}_i \tag{4.7}$$

Using ${}^{\mathcal{I}}\boldsymbol{v}_i = ({}^{\mathcal{I}}x_i, {}^{\mathcal{I}}y_i, {}^{\mathcal{I}}z_i)^T$ and ${}^{\mathcal{C}_R}\boldsymbol{v}_i = ({}^{\mathcal{C}_R}x_i, {}^{\mathcal{C}_R}y_i, {}^{\mathcal{C}_R}z_i)^T$, according to Lobo and Dias [2007], this quaternion product can be expressed in matrix form, being equivalent to

$$\max \mathring{\mathbf{q}}^T \boldsymbol{N} \mathring{\mathbf{q}} \tag{4.8}$$

where

$$\boldsymbol{N} = \sum_{i=1}^{n} {}^{\mathcal{I}}\mathbf{V}_i^T \cdot {}^{\mathcal{C}_R}\mathbf{V}_i$$

with

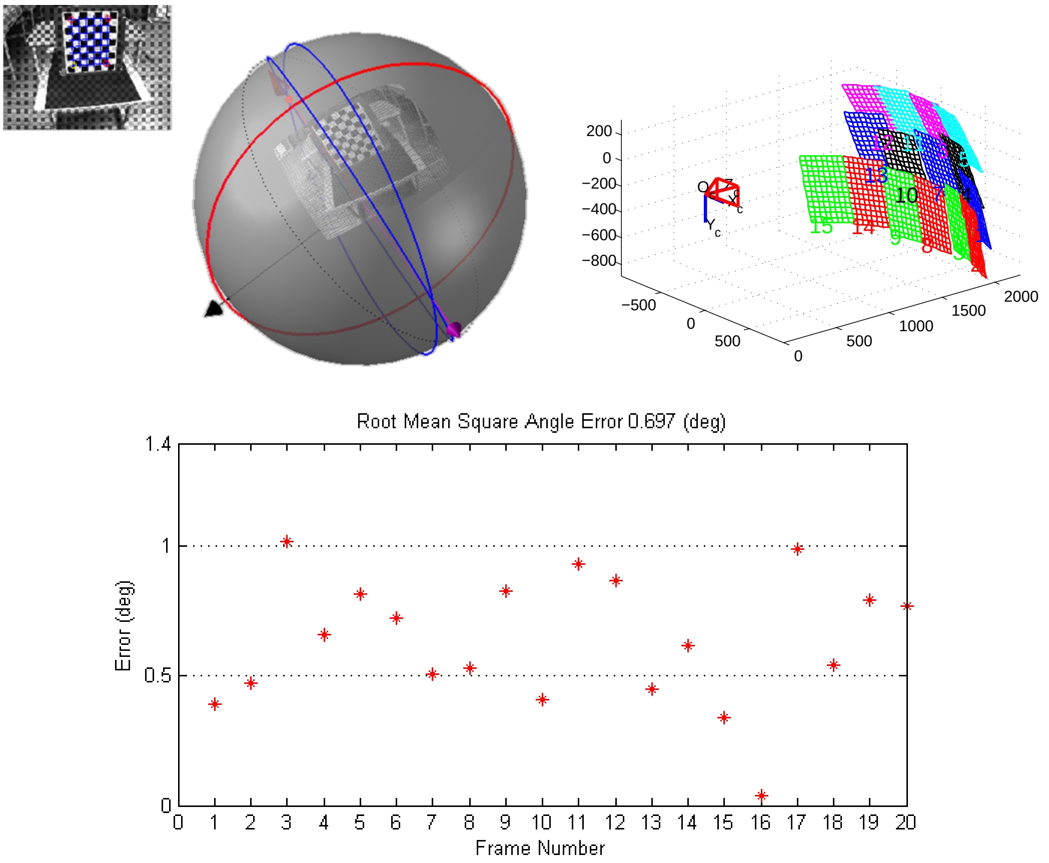


Figure 4.10: Outcome of visuoinertial calibration of the IMPEP V2.0 platform. Top left: unit sphere projection with vanishing point and reprojected verticals from rotation calibration. Top right: reconstructed target positions relative to the camera. Bottom: reprojection alignment errors for verticals in each of the 20 frames used for camera calibration and rotation estimation.

$$\mathbf{V}_i = \begin{bmatrix} 0 & -{}^{C_R}x_i & -{}^{C_R}y_i & -{}^{C_R}z_i \\ {}^{C_R}x_i & 0 & -{}^{C_R}z_i & {}^{C_R}y_i \\ {}^{C_R}y_i & {}^{C_R}z_i & 0 & -{}^{C_R}x_i \\ {}^{C_R}z_i & -{}^{C_R}y_i & {}^{C_R}x_i & 0 \end{bmatrix}$$

This boresight static approach can be easily performed, not requiring any additional equipment, apart from the chessboard target, obtained using a standard printer, already used for camera calibration [Lobo and Dias 2007]. Fig. 4.9 provides a summary of required steps to perform calibration of rotation between camera and IMU as described in Lobo and Dias [2007].

Visuoinertial calibration was performed using the InerVis toolbox [Lobo 2006], that adds on to the Camera Calibration Toolbox by Bouguet [2006]. An overview of the outcome of the visuoinertial calibration process of the IMPEP V.2 platform is shown in Fig. 4.10.

Binaural system calibration

As can be seen on the BP in Fig. 2.12, calibration of the binaural system involves the characterisation of the families of normal distributions $P(\tau|S_C\,O_C\,\theta_{\max})$ and $P(\Delta L(f_c^k)|\tau\,S_C\,O_C\,C) \approx P(\Delta L(f_c^k)|S_C\,O_C\,C)$ through descriptive statistical learning of their central tendency and statistical variability. This is done in an equivalent manner as with commonly used head-related transfer function (HRTF) calibration processes (see, for example, [Calamia 1998]) and is described in the following paragraphs.

A set M_c of n-dimensional measurement vectors such as defined on chapter 2 is collected per cell $c \in \mathcal{C}$. The full set of collected measurement vectors for all cells in auditory sensor space $\mathcal{Y}$ is expressed as $M = \bigcup M_c$. Denoting $M_{\bar{c}} = M \setminus M_c$ as the set of measurements for all cells other than c, the statistical characterisation process of each family of distributions is effected for each cell c through

$$P(\tau|[S_c = 1]\,O_c\,\theta_{\max}) \equiv \mathcal{N}(\tau, \mu_\tau(M_c), \sigma_\tau(M_c)) \tag{4.9a}$$

$$P(\tau|[S_c = 0]\,O_c\,\theta_{\max}) \equiv \mathcal{N}(\tau, \mu_\tau(M_{\bar{c}}), \sigma_\tau(M_{\bar{c}})) \tag{4.9b}$$

$$\begin{aligned} &P(\Delta L(f_c^k)|[S_c = 1]\,O_c\,c) \equiv \\ &\quad \mathcal{N}(\Delta L(f_c^k), \mu_{\Delta L(f_c^k)}(M_c), \sigma_{\Delta L(f_c^k)}(M_c)) \end{aligned} \tag{4.9c}$$

$$\begin{aligned} &P(\Delta L(f_c^k)|[S_c = 0]\,O_c\,c) \equiv \\ &\quad \mathcal{N}(\Delta L(f_c^k), \mu_{\Delta L(f_c^k)}(M_{\bar{c}}), \sigma_{\Delta L(f_c^k)}(M_{\bar{c}})) \end{aligned} \tag{4.9d}$$

Auditory calibration is performed by presenting a broadband audio stimulus through a loudspeaker positioned in well-known spatial coordinates corresponding to the geometric centre of each cell $c \in \mathcal{C}$ so as to sample space according to the auditory sensor space $\mathcal{Y}$.

The acquisition method may be simplified by a factor of 4 by taking into account the spatial redundancies of auditory sensing, namely the symmetry enforced by the back-to-front ambiguity and the left-to-right antisymmetry for both ITDs and ILDs, to reduce calibration space to the front-left quadrant.

A further simplification of the procedure consists in positioning the loudspeaker, for each of the N_d considered distances from the binaural system, precisely in front of the active perception head (i.e. $(\theta, \phi) = (0, 0)$) and to *rotate the active head* so that

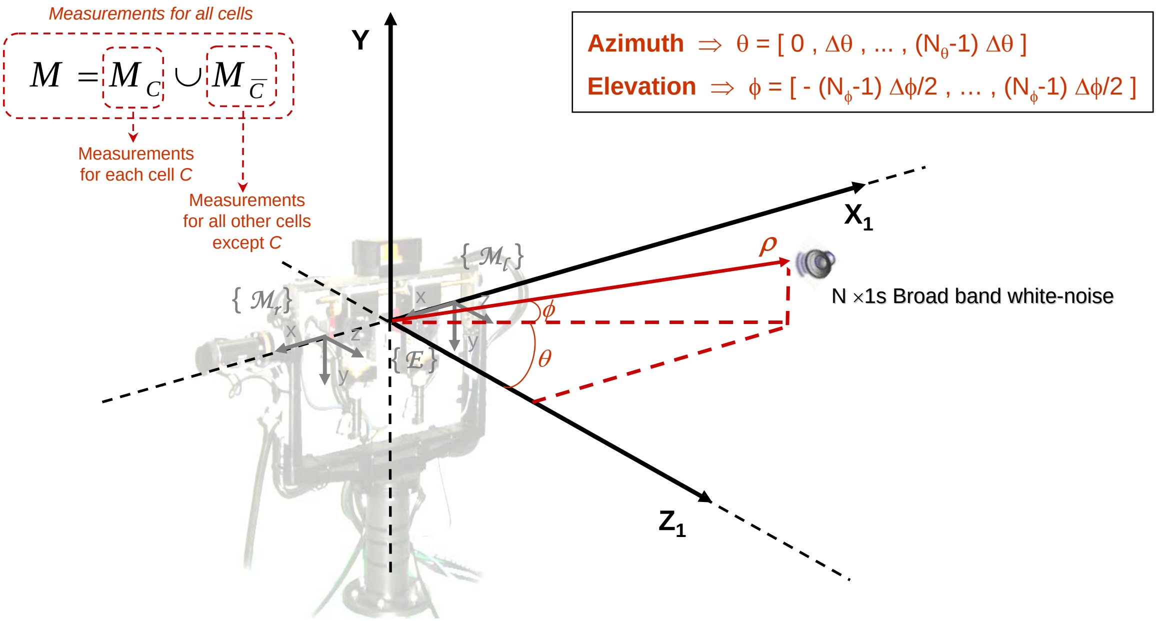

Measurements for all cells
$M = M_C \cup M_{\bar{C}}$
Measurements for each cell C
Measurements for all other cells except C
Y
X_1
Z_1
ρ
ϕ
θ
N ×1s Broad band white-noise
$\{\mathcal{M}_l\}$
$\{\mathcal{M}_r\}$
$\{\mathcal{E}\}$
x
y
z
Azimuth ⇒ $\theta = [\,0\,, \Delta\theta\,, \ldots\,, (N_\theta-1)\,\Delta\theta\,]$
Elevation ⇒ $\phi = [\,-(N_\phi-1)\,\Delta\phi/2\,, \ldots\,, (N_\phi-1)\,\Delta\phi/2\,]$

Figure 4.11: Experimental setup for the binaural system calibration procedure.

the whole range of azimuths and elevations of the auditory sensor space is covered. This replaces the several minutes taken to reposition the loudspeaker by hand (now only happening N_d times) by a few seconds of head motions for each cell. The full procedure is depicted in Fig. 4.11.

4.3 Results and Conclusions

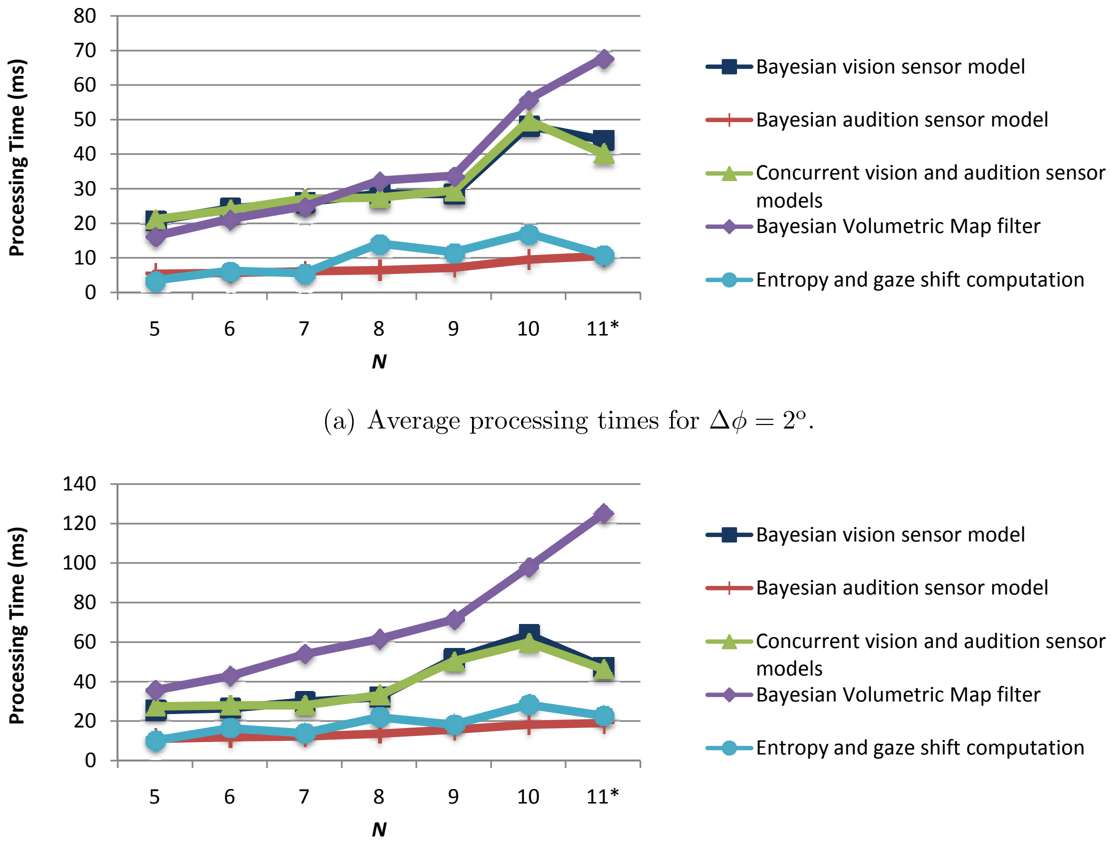
The real-time implementation of all the processes of the framework was subjected to performance testing for each individual module. Processing times and rates for the sensory systems are as follows:

- **Stereovision unit** 15 Hz, including image grabbing and preprocessing (using CPU), stereovision processing itself (i.e. disparity and confidence map generation, using GPU), and postprocessing and numerical conditioning (using CPU).

- **Binaural processing unit** Maximum rate of 40 Hz and 20 to 70 ms latency (using CPU) for 44 KHz, 16-bit audio, with 16 frequency channels and 50 ms buffer for cue computation.

- **Inertial processing unit** 100 Hz using GPU.

Processing times for the BVM modules are shown in Fig. 4.12. As can be seen, the full active exploration system runs from 6 to 10 Hz, depending on system parameters.



(a) Average processing times for $\Delta\phi = 2^o$.

(b) Average processing times for $\Delta\phi = 1^o$.

Figure 4.12: BVM framework average processing times. Both graphs are for $\Delta\theta = 1^o$, and show the average of processing times in ms for each activity depicted on Fig. 4.5, taken for a random set of 500 runs of each module in the processing of 5 dynamic real-world scenarios, with sensory horopter occupation varying roughly from 10 to 40% (although with no apparent effect on performance). These times are plotted against the number N of divisions in distance, which is the most crucial of system parameters (for $N > 11$, the GPU resources become depleted, and for $N < 5$ resolution arguably becomes unsatisfactory), and for two different reasonable resolutions in ϕ. Note that BVM filter performance degrades approximately exponentially with increasing resolution in distance, while the performance of all other activities degrades approximately linearly — the sole exception is the vision sensor model for $N = 11$, where it actually improves its performance. The reason for this is that the ratio of the effect of the influence of resolution on CUDA grid size *vs* the effect of the influence of resolution on the number of atomic operations required is reversed. (The * denotes that for $N = 11$ the block size is smaller for the BVM filter CUDA implementation — refer to main text for further details.)

This is ensured by forcing the main BVM thread to pause for each time-step when no visual measurement is available (e.g. during 40 ms for $N = 10, \Delta\phi = 2^{\mathrm{o}}$ — see Fig. 4.5). 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.

Running times for the Bayesian Volumetric Map update process decreased for each processing cycle from 5 to 30 minutes of serial processing on a Pentium Core 2 Quad CPU at 2.40 GHz, depending on BVM parameters, to a corresponding few hundredths of a second to a tenth of a second of parallel computing on an NVIDIA 9800 GTX graphics card, thus yielding a $18,000$ to $30,000$-times faster performance.

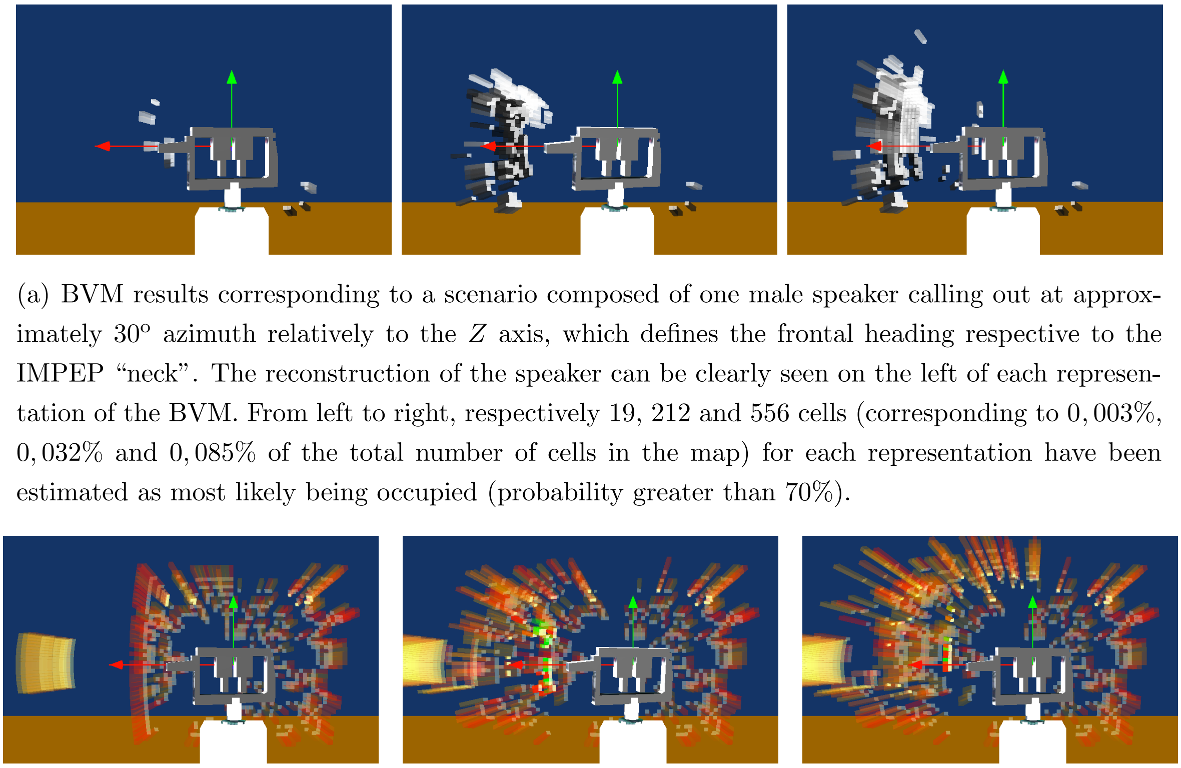
An exemplification of the active exploration algorithm performing in real-time is presented on Fig 4.13. Results consist of offline rendering of BVMs computed online at time-instants in which gaze shifts took place, together with a corresponding representation of relevant values of the entropy-based variable U_C for each BVM cell.

Online results of processing three disparate scenarios testing different aspects of the full system are presented on Figs. 4.14, 4.15 and 4.16. Scenes consisting of one or more speakers talking in a cluttered lab are observed by the IMPEP active perception system and processed online by the BVM Bayesian filter, using entropy-based active exploration as described earlier, in order to scan the surrounding environment.

More specifically, on Fig. 4.14 a 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. Figs. 4.15 and 4.16 show 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 and three-speaker scenario, respectively, since in both cases the BVM successively generates a reconstruction of each of the speakers.

These results show a projection of the log-spherical configuration onto Euclidean space of a volume 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 [Cutting and Vishton 1995]) and the evolution of the exploration process through time.

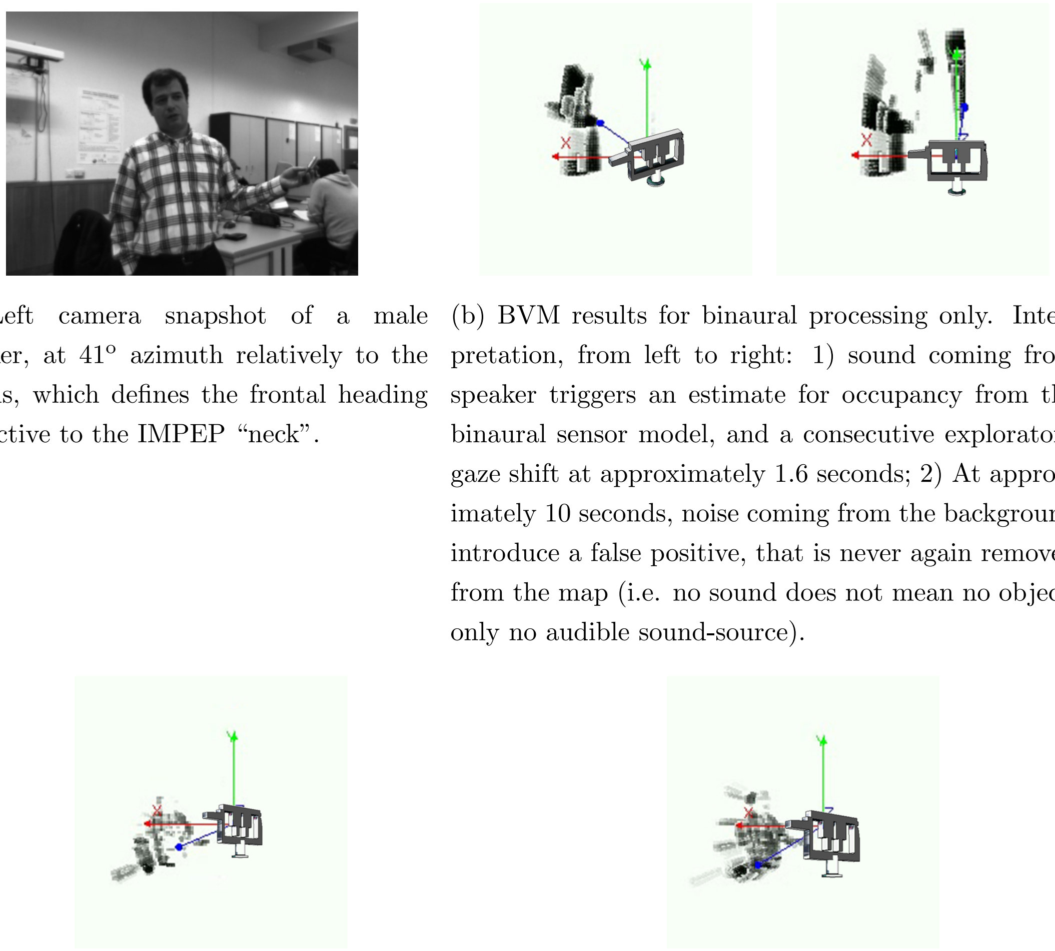
The active exploration algorithm thus successfully drives the IMPEP-BVM framework to explore areas of the environment mapped with high uncertainty in real-time, with an intelligent heuristic that minimises the effects of local minima by attending to



(a) BVM results corresponding to a scenario composed of one male speaker calling out at approximately 30º azimuth relatively to the Z axis, which defines the frontal heading respective to the IMPEP "neck". The reconstruction of the speaker can be clearly seen on the left of each representation of the BVM. From left to right, respectively 19, 212 and 556 cells (corresponding to 0, 003%, 0, 032% and 0, 085% of the total number of cells in the map) for each representation have been estimated as most likely being occupied (probability greater than 70%).

(b) Relevant values for entropy-based variable U_C corresponding to each of the time-instants in (a). Represented values range from .5 to 1, depicted using a smoothly gradated red-to-green colour-code (red corresponds to lower values, green corresponds to higher values). Chronologically ordered interpretation of these results goes as follows: at first, relevant cells have their relative importance for sensory exploration scattered throughout the visible area, and there is a separate light yellow region on the left corresponding to an auditory object (i.e the speaker) that becomes the focus of interest (816 cells — 0.126% of the total number of cells in the map — were considered as minimally relevant); then, at the boundaries of the speaker's silhouette, bright green cells show high relevance of this area for exploration, which then becomes the next focus of interest (1, 710 cells — 0.264% — were considered as minimally relevant); finally, after a few cycles of BVM processing, uncertainty lowers, which clearly shows as the number of green cells diminishes (1, 901 cells — 0.293% — were considered as minimally relevant).

Figure 4.13: Results corresponding, from left to right, to time-instants in which gaze shifts were generated, 17.080 s, 26.664 s and 36.411 s, respectively, for the real-time prototype for multimodal perception of 3D structure and motion using the BVM, exemplifying the use of the entropy-based variable U_C to elicit gaze shifts, using the active exploration heuristics described in the main text, in order to scan the surrounding environment. A scene consisting of a male speaker talking in a cluttered lab is observed by the IMPEP active perception system and processed online by the BVM Bayesian filter. An oriented 3D avatar of the IMPEP perception system depicted in each map denotes the current gaze orientation. All results depict frontal views, with Z pointing outward. The parameters for the BVM are as follows: $N = 10$, $\rho_{Min} = 1000\,\text{mm}$ and $\rho_{Max} = 2500\,\text{mm}$, $\theta \in [-180^o, 180^o]$, with $\Delta\theta = 1^o$, and $\phi \in [-90^o, 90^o]$, with $\Delta\phi = 1^o$, corresponding to $10 \times 360 \times 180 = 648,000$ cells, approximately delimiting the so-called "personal space" (the zone immediately surrounding the observer's head, generally within arm's reach and slightly beyond, within 2 m range [Cutting and Vishton 1995]).



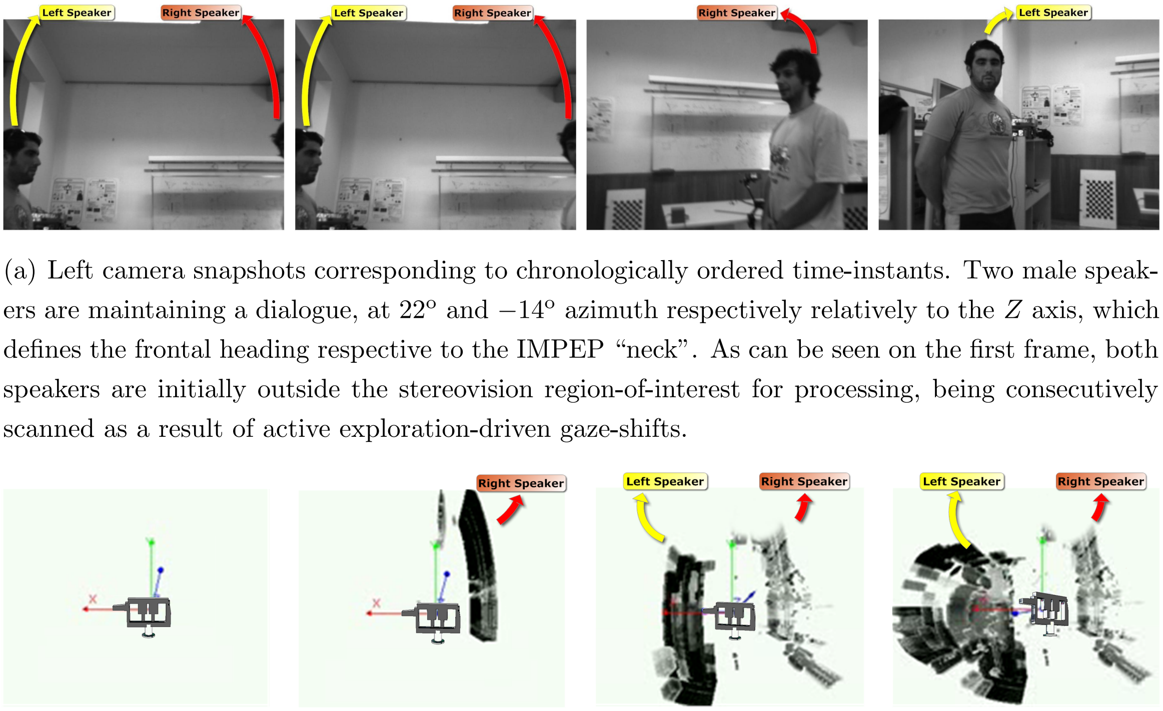
(a) Left camera snapshot of a male speaker, at 41° azimuth relatively to the Z axis, which defines the frontal heading respective to the IMPEP "neck".

(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.

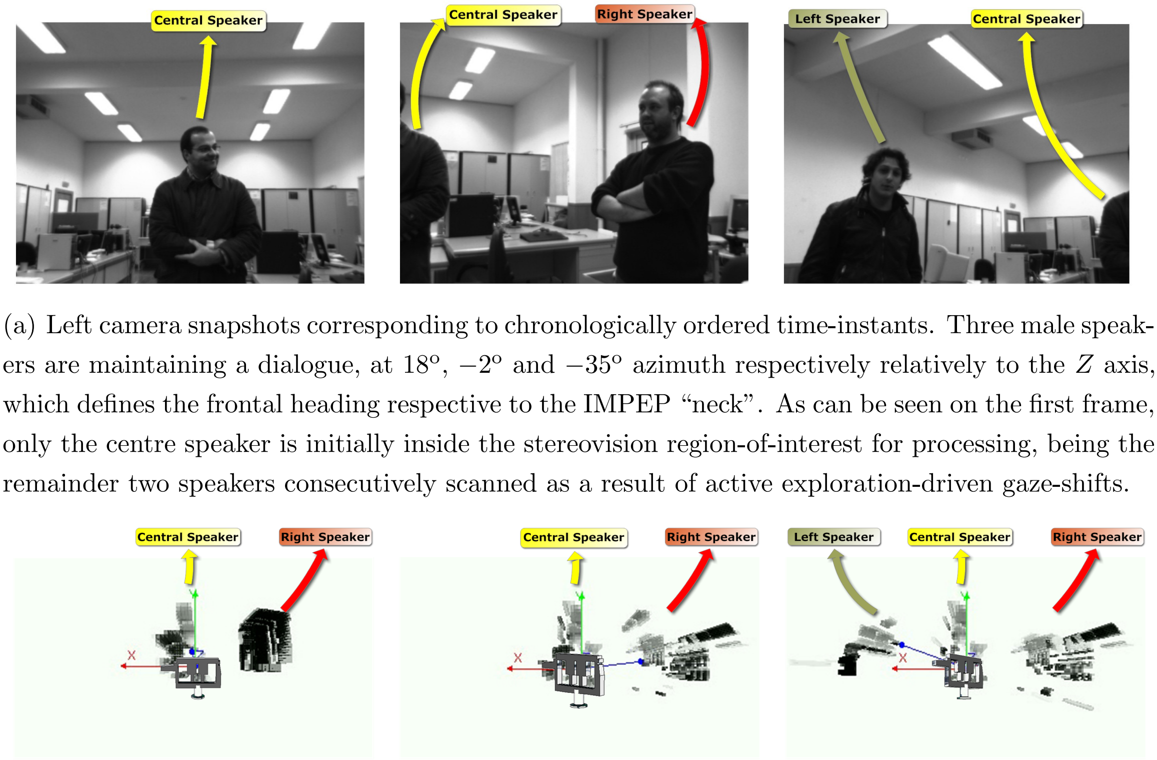
Figure 4.14: Results for the real-time prototype for multimodal perception of 3D structure and motion using the BVM — three reenactments (binaural sensing only, stereovision sensing only and visuoauditory sensing) of a single speaker scenario. A scene consisting of a male speaker talking in a cluttered 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 blue 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 $\rho_{Max} = 2500\,\text{mm}$, $\theta \in [-180^o, 180^o]$, with $\Delta\theta = 1^o$, and $\phi \in [-90^o, 90^o]$, with $\Delta\phi = 2^o$, 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 [Cutting and Vishton 1995]).



(a) Left camera snapshots corresponding to chronologically ordered time-instants. Two male speakers are maintaining a dialogue, at 22° and −14° azimuth respectively relatively to the Z axis, which defines the frontal heading respective to the IMPEP "neck". As can be seen on the first frame, both speakers are initially outside the stereovision region-of-interest for processing, being consecutively scanned as a result of active exploration-driven gaze-shifts.

(b) BVM results corresponding to each of the snapshots in (a). Interpretation, from left to right (chronological evolution): 1) initial non-informative map; 2) sound coming from the speaker on the right triggers an estimate for occupancy from the binaural sensor model, and a consecutive exploratory gaze shift; 3) a few frames from the stereovision system trigger further evidence accumulation for occupancy by the vision sensor model at the gaze direction site, fusing readings from both sensory systems — higher spatial resolution from vision carves out the right speaker's silhouette from the first rough estimate from audition —, while sound coming from the speaker on the left triggers an estimate for occupancy from the binaural sensor model, and a consecutive exploratory gaze shift in the speaker's direction; 4) after turning to new gaze direction site, the stereovision system triggers further evidence accumulation for occupancy at the left speaker's location, fusing readings from both sensory modalities.

Figure 4.15: Results for the real-time prototype for multimodal perception of 3D structure and motion using the BVM — two speakers scenario. A scene consisting of two male speakers talking to each other (left and right speakers both pinpointed for clarity) 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. All parameters and labelling are as in Fig. 4.14, unless where otherwise noted.



(a) Left camera snapshots corresponding to chronologically ordered time-instants. Three male speakers are maintaining a dialogue, at 18°, -2° and -35° azimuth respectively relatively to the Z axis, which defines the frontal heading respective to the IMPEP "neck". As can be seen on the first frame, only the centre speaker is initially inside the stereovision region-of-interest for processing, being the remainder two speakers consecutively scanned as a result of active exploration-driven gaze-shifts.

(b) BVM results corresponding to each of the snapshots in (a). Interpretation, from left to right (chronological evolution): 1) while central speaker within sight, sound coming from the speaker on the right (out of sight) triggers an estimate for occupancy from the binaural sensor model, and a consecutive exploratory gaze shift; 2) a few frames from the stereovision system trigger further evidence accumulation for occupancy by the vision sensor model at the gaze direction site, fusing readings from both sensory systems; 3) sound coming from the speaker on the left triggers an estimate for occupancy from the binaural sensor model, and a consecutive exploratory gaze shift in the speaker's direction; after turning to new gaze direction site, the stereovision system triggers further evidence accumulation for occupancy at the left speaker's location, fusing readings from both sensory modalities.

Figure 4.16: Results for the real-time prototype for multimodal perception of 3D structure and motion using the BVM — three speakers scenario. A scene consisting of three male speakers talking to each other 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. All parameters and labelling are as in Figs. 4.14 and 4.15, unless where otherwise noted.

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 [Carpenter 2004, Caspi et al. 2004], the overall rates of 6 to 10 Hz achieved with our CUDA implementation, in our opinion, back up the claim that our system does, in fact, achieve satisfactory real-time performance.

Chapter 5

A Bayesian Hierarchical Framework for Multimodal Active Perception

5.1 Motivations

Kopp and Gärdenfors [2002] posit that the capacity of attention is a *minimal criterion of intentionality* for robots. Since humans are prevalently social beings, their attentional system is inherently socially driven; this becomes particularly important when considering human-machine interaction, where robots are expected to engage with humans while displaying attentional behaviours that resemble those of their interlocutors [Shic and Scassellati 2007].

In any case, even when dealing with unknown environments with no social intent, humans use their attentional system to its full — a random exploratory strategy alone would not take into account potential dangers that our evolution has imprinted in our prior knowledge to most probably be due to predators or caused by competitors of our own species.

Given our bioinspired motivations presented in the introductory chapter, we set off to develop a hierarchical artificial active perception system that follows human-like bottom-up driven behaviours based on vision, audition and proprioception, using the framework presented on chapters 2 and 4. In the process, we will demonstrate the following properties which are intrinsic to the framework: *emergence*, *scalability* and *adaptivity*.

5.2 A Bayesian Hierarchy as a Model of Active Visuoauditory Perception

5.2.1 Background and definitions

A saccade is a fast movement of an eye, head or other part of an animal's body or device. For example, eye saccades are quick, simultaneous movements of both eyes in the same direction. Saccades serve several purposes, such as a mechanism for fixation or rapid eye movement [Ferreira and Castelo-Branco 2007].

Humans and other animals do not look at a scene in a steady way. Instead, sensors are directed to interesting parts of the scene, so as to build up a mental map of the surrounding environment. One reason for saccades is to move the senses so that redundant evidence can be accumulated about a scene (i.e. active perception), lowering the overall uncertainty of individual sensor measurements and using limited-scope sensorial resources more efficiently.

Visual saccades, the most thoroughly investigated type of saccade, are measured or investigated in four ways (which can be generalised to multisensory-driven saccades) [Rommelsea et al. 2008]:

- In a visually guided saccade, an observer performs a *gaze shift* towards a visual onset, or stimulus. This is typically included as a baseline when measuring other types of saccades.

- In an antisaccade, an observer moves eyes away from the visual onset. They are more delayed than visually guided saccades, and observers often make erroneous saccades in the wrong direction. A successful antisaccade requires inhibiting a reflexive saccade to the onset location, and voluntarily moving the eye in the other direction.

- In a memory guided saccade, an observer shifts its gaze towards a remembered point, with no sensory onset involved.

- In smooth pursuit eye movements, an observer tracks a small object moving with a constant slow speed. They emphasise basic eye control, not cognitive processes.

See Fig. 5.1 for an overview of visual saccadic and smooth pursuit circuits of the Macaque monkey brain, which have analogous counterparts in the human brain.

Sensory guided and memory guided saccades involve *gaze computation*, the object of the models presented herewith, followed by *gaze control*, which translates desired

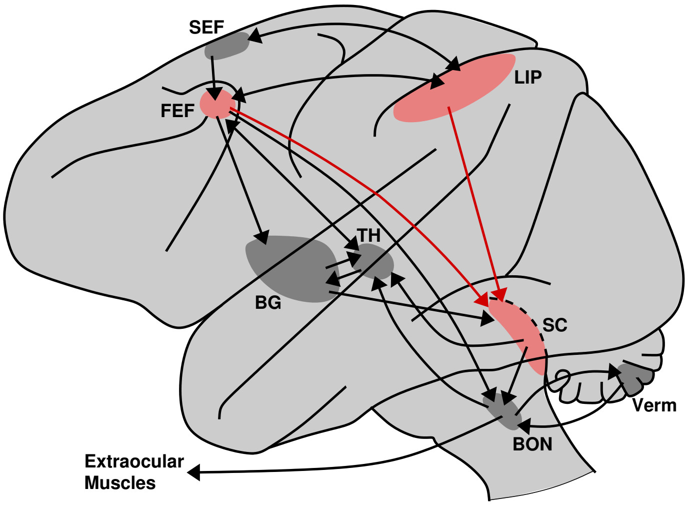


Figure 5.1: Premotor and motor circuitry shared by saccade and smooth pursuit movement (Macaque monkey). BG: basal ganglia, BON: brainstem oculomotor nuclei, FEF: frontal eye fields, LIP: lateral bank of the intraparietal sulcus, SC: superior colliculus, SEF: supplementary eye fields, TH: thalamus, Verm: cerebellar vermis. In red: regions using retinotopic reference frames to encode visual, memory and motor activity. Reproduced with kind permission from Colas et al. [2009].

fixation points to sequences of commands to the eye and head (i.e. motor commands) and is beyond the scope of this text. Gaze computation is typically broken up into two phases: an *attention model* that identifies relevant features in the scene, selects one of these features and maintains focus on it, and a *gaze policy*, that operates over the feature map to determine the actual fixation point [Shic and Scassellati 2007, Kopp and Gärdenfors 2002].

There are many ways that can be used to classify an attention system according to its various aspects. In a subject's point of view, the subject can actually switch the gaze fixation point to the point being attended to (i.e., *overt attention*). On the other hand, it can also shift the attentional processing without any a fixation shift or motor action (i.e., *covert attention*). We will be focusing our attention on the former.

On the other hand, in order that behaviourally relevant perceptual information is appropriately selected, efficient mechanisms must be in place. Two major attentional mechanisms are known to control this selection process [Parkhurst et al. 2002]. First, bottom-up attentional selection is a fast, and often compulsory, stimulus-driven mechanism (related to the so-called *exogenous attention*). There is now clear evidence indicating that attention can be captured under the right stimulus conditions. For example, highly salient feature singletons or abrupt onsets of new perceptual objects automatically attract attention (pop-up effects). The other mechanism, top-down attentional selection, is a slower, goal-directed mechanism where the observer's expectations or intentions influence the allocation of attention (related to the so-called *endogenous attention*). Observers can volitionally select regions of space or individual objects to attend. The degree to which these two mechanisms play a role in determining attentional selection under natural viewing conditions has been for a long time under debate [Parkhurst et al. 2002].

A great deal of research has been dedicated to developing models of visual attention in the past few years. These computational models are just rough approximations to the human visual attention system and typically operate by identifying, within an incoming visual stream, spatial points of interest. This computational formulation of perceptual attention is very limiting, in terms of the capabilities and complexities of the biological reality [Shic and Scassellati 2007]. These models serve to reduce the scene to several points of particular interest, and to emulate the scan-path behaviour of human subjects. In this fashion, it is possible to control the combinatorial explosion that results from the consideration of all possible image relationships and provide a naturalistic interface to behaviours such as joint attention [Shic and Scassellati 2007].

However, even in visual animals *multisensory* stimuli (e.g. visual, auditory or tactile) elicit gaze shifts to aid visual perception of stimuli. Such gaze shifts can either be top-down attention driven (e.g. visual search) or they can be reflex movements triggered by unexpected changes in the surroundings triggered by the collective result of multimodal perception [Koene et al. 2007].

Several representative models addressing most of these issues will be briefly reviewed in the following lines.

One of the most popular computational models serving as a basis for robotic implementations of visual attention is the model by Itti et al. [1998]. This model has roots at least as far back as [Niebur et al. 1995] and its most recent developments are described in [Carmi and Itti 2006].

Itti et al.'s model is a feed-forward bottom-up computational model of visual attention, employing, at its most basic level, decompositions into purely preattentive features. This gives advantages in both speed and transparency. It is a model that is not only simple but also rigorously and specifically defined, a strong advantage for implementation, extension, and reproducibility of results [Shic and Scassellati 2007]. The model extracts the preattentive modalities of color, intensity, and orientation from an image. These modalities are assembled into a multiscale representation using Gaussian and Laplacian pyramids. Within each modality, centre-surround operators are applied in order to generate multiscale feature maps. An approximation to lateral inhibition is then employed to transform these multiscale feature maps into conspicuity maps, which represent the saliency of each modality. Finally, conspicuity maps are linearly combined to determine the saliency of the scene. Although this model did not originally attend to visual motion, known to be a major modality in visual attention, it has been extended to include it in later work, such as [Shic and Scassellati 2007].

Parkhurst, Law, and Niebur [2002] show that the saliency maps of images, as computed by the Itti model, display higher values in locations fixated upon by human subjects than would have been expected by chance alone. The fact that the saliency maps generated by the same computational attention model can be correlated to approximate probability density maps of humans is shown by Ouerhani et al. [2004]. The model by Itti et al. is not uncontroversial, as can be seen in the completely different evaluations by Parkhurst et al. [2002], who generally validate the model, and Turano, Geruschat, and Baker [2003], who attempt to detract from it by claiming that the predicted gaze locations are no better than random, or Tatler et al. [2005a], who claim that the model is not scale or rotation invariant, thus questioning the appropriateness of using it as the basis of computational object recognition systems[1]. In any case, Itti and coworkers have shown that interesting objects seem to be visually salient, indicating that selecting interesting objects in the scene is largely constrained by low-level visual properties rather than solely determined by higher cognitive processes [Elazary and Itti 2008]. Shic and Scassellati [2007] build upon the Itti model to apply a framework, based on dimensionality-reduction over the features of human gaze trajectories, that can simultaneously be used for both optimising a particular computational model of visual attention and for evaluating its performance in terms of similarity to human behaviour.

Alternative computational models of visual attention both with and without motion besides the model presented above exist, such as the work of Tsotsos et al. [1995] or Breazeal and Scassellati [1999] and many others.

The gaze computation process takes, as an input, the saliency map, and returns, as an output, a point of fixation. One of the simplest gaze policies that can by employed is to simply index the location in the saliency map corresponding to the highest peak [Shic and Scassellati 2007].

On the other hand, regarding the *temporal* dimension of attention, a commonly used complementary model is the Inhibition of Return (IoR) mechanism [Niebur, Itti, and Koch 1995]. The IoR, in simple terms, is the mechanism where the saccade gener-

[1] For a deeper insight, please refer to [Shic and Scassellati 2007].

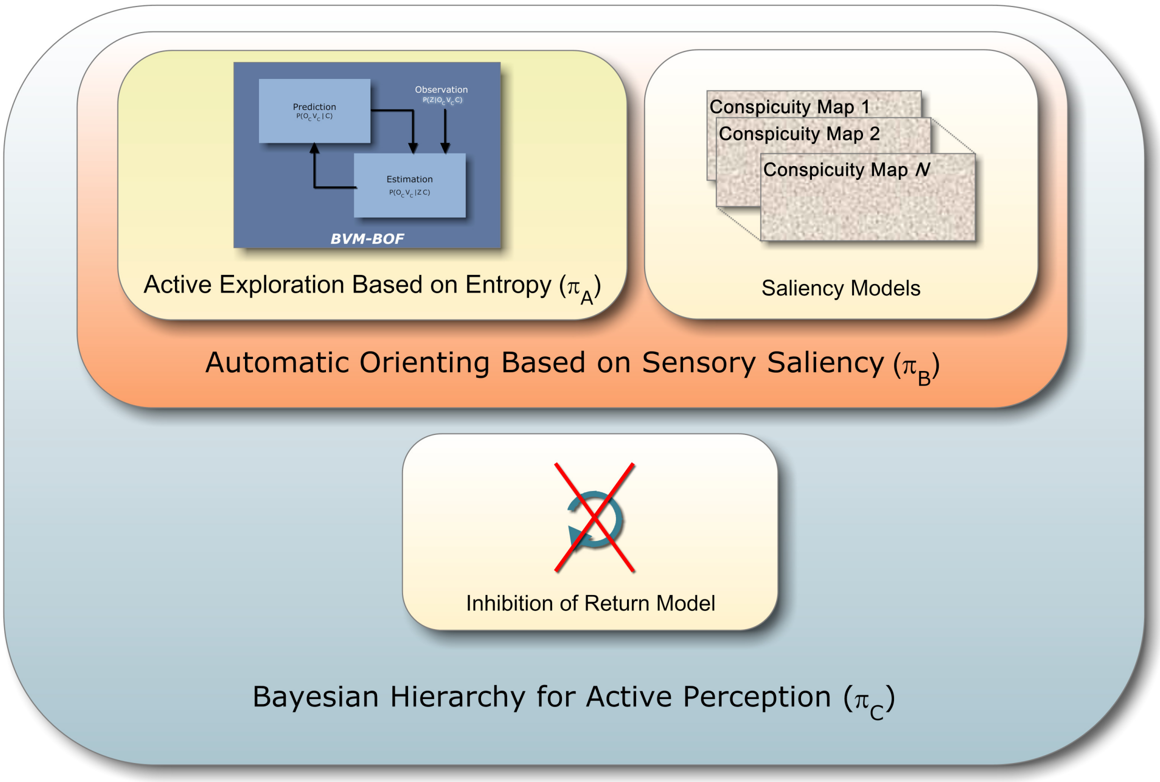

Prediction
P(O_C V_C | C)
Observation
P(Z|O_C V_C C)
Estimation
P(O_C V_C |Z C)
BVM-BOF
Active Exploration Based on Entropy (π_A)
Conspicuity Map 1
Conspicuity Map 2
Conspicuity Map N
Saliency Models
Automatic Orienting Based on Sensory Saliency (π_B)
Inhibition of Return Model
Bayesian Hierarchy for Active Perception (π_C)

Figure 5.2: Active perception model hierarchy.

ating system in the brain avoids fixation sites which have just been a focus of attention, therefore preventing deadlocks. Recently, a more complex model has been devised, using Bayesian surprise as a factor related to the attentional changes in the time domain, by Itti and Baldi [2009].

5.2.2 Models

To achieve our goal of designing Bayesian models for visuoauditory-driven saccade generation following human active perception behaviours, a hierarchical framework, inspired on what was proposed by Colas, Flacher, Tanner, Bessière, and Girard [2009], has been developed and is presented in the following text.

We will specify three decision models[2]: π_A, that implements entropy-based active exploration based on the BVM (chapter 2) and the heuristics presented on chapter 4, π_B, that uses entropy and saliency together for active perception, and finally π_C which adds a simple Inhibition of Return mechanism based on the fixation point of the previous time-step. In other words, each model π_k incorporates its predecessor π_{k-1} through

[2] Refer to Appendix A for a formal definition of π_k within the Bayesian Programming formalism.

Program

Description

Specification

Relevant variables:

O_C^t: binary value describing the occupancy of cell C at time t, $[O_C = 1]$ if cell C is occupied by an object, $[O_C = 0]$ otherwise (related variable: $O^t = \bigwedge_C O_C^t$);

V_C^t: velocity of cell C at time t, discretised into $n+1$ possible cases $\in \mathcal{V} \equiv \{v_0, \cdots, v_n\}$ (related variable: $V^t = \bigwedge_C V_C^t$);

$G^t \equiv (\log_b \rho_{\max}, \theta_{\max}, \phi_{\max}) \in \mathcal{Y}$: fixation point for next gaze-shift, computed at time t, with related variable $G = \bigwedge_{t \in [1, t_{max}]} G^t$;

$U_C^t \equiv f(V^{1 \to t}, O^{1 \to t}) \in [0,1]$: joint entropy gradient-based variable at time t, with related variables $U^t = \bigwedge_C U_C^t$ and $U = \bigwedge_{t \in [1, t_{max}]} U^t$

(see equation (4.3): close to 1 when uncertainty is high and C is a frontier cell, and $U_C \to 0$ when uncertainty is low or C is not a frontier cell).

Decomposition:

$$P(G\,U|\pi_A) = \prod_{t=1}^{t_{max}} \left[P(G^t|\pi_A) \prod_C P(U_C^t|G^t\,\pi_A) \right]$$

Parametric forms:

$P(G^t|\pi_A)$: uniform prior;

$P(U_C^t|G^t\,\pi_A)$: is a beta distribution $\mathcal{B}(\alpha_U, \beta_U)$ for $[G^t = C]$ that expresses that, for a given point of fixation proposal for the next gaze shift, U_C^t is more likely near 1, or a uniform distribution on U_C^t for $[G^t \neq C]$.

Identification:

Empirical values for free parameters α_U and β_U.

Questions:

$P(G^t|V^{1 \to t}\,O^{1 \to t}\,\pi_A) = P(G^t|U^t\,\pi_A)$

Figure 5.3: Bayesian Program for entropy-based active exploration model π_A.

Bayesian fusion, therefore constituting a model hierarchy — see Fig. 5.2.

The hierarchy is extensible in such a way that other properties characterised by additional random variables and corresponding probabilities might be represented, other than the already implemented occupancy and local motion properties of the BVM, by augmenting the hierarchy of operators through Bayesian subprogramming [Bessière et al. 2008, Lebeltel 1999]. This ensures that the framework is *scalable*. On the other hand, the combination of these strategies to produce a coherent behaviour ensures that the framework is *emergent*.

Furthermore, each model will infer a probability distribution on the next point of fixation for the next desired gaze shift represented by a random variable $G^t \in \mathcal{Y}$ at each time $t \in [1, t_{max}] : P(G^t|V^{1 \to t}\,O^{1 \to t}\,\pi_k)$, where $V^{1 \to t} = \bigwedge_{t \in [1,t_{max}]} \bigwedge_C V_C^t$ and $O^{1 \to t} = \bigwedge_{t \in [1,t_{max}]} \bigwedge_C O_C^t$ represent the conjunction of BVM local motion and occupancy estimate states for all cells $C \in \mathcal{Y}$ at from system startup to time t.

Program

Description

Specification

Relevant variables:

O_C^t: binary value describing the occupancy of cell C at time t, $[O_C = 1]$ if cell C is occupied by an object, $[O_C = 0]$ otherwise (related variables: $O = \bigwedge_{t\in[1,t_{max}]} O_C^t$ and $O^t = \bigwedge_C O_C^t$);

V_C^t: velocity of cell C at time t, discretised into $n+1$ possible cases $\in \mathcal{V} \equiv \{v_0, \cdots, v_n\}$ (related variables: $V = \bigwedge_{t\in[1,t_{max}]} V_C^t$ and $V^t = \bigwedge_C V_C^t$);

$G^t \equiv (\log_b \rho_{\max}, \theta_{\max}, \phi_{\max}) \in \mathcal{Y}$: fixation point for next gaze-shift, computed at time t, with related variable $G = \bigwedge_{t\in[1,t_{max}]} G^t$;

$S_C^{i,t}$: binary value describing the ith of N sensory saliency properties of cell C at time t, $[S_C^{i,t} = 0]$ when non-salient and $[S_C^{i,t} = 1]$ when salient (related variables: $S^i = \bigwedge_{t\in[1,t_{max}]} S^{i,t}$, $S^t = \bigwedge_{i=1}^{N} S^{i,t}$ and $S = \bigwedge_{i=1}^{N} S^i$);

$Z_j^{i,t} \in \mathcal{Z}$: sensor measurements at time t ($j = 1..M_i$ total independent measurements for each saliency property at time t, $S^{i,t}$) (related variables: $Z^i = \bigwedge_{t\in[0,t_{max}]} Z^{i,t}$ and $Z = \bigwedge_{i=1}^{N} Z^i$);

$Q_C^{i,t} = P([S_C^{i,t} = 1] | Z_j^{i,t}\, C) \in [0,1]$: probability of a perceptually salient object occupying cell C (related variables: $Q^i = \bigwedge_{t\in[1,t_{max}]} Q^{i,t}$, $Q^t = \bigwedge_{i=1}^{N} Q^{i,t}$ and $Q = \bigwedge_{i=1}^{N} Q^i$).

Decomposition:

$$P(G\,Q|\pi_B) = \prod_{t=1}^{t_{max}} \left\{ P(G^t|\pi_B) \prod_C \left[\prod_{i=1}^{N} P(Q_C^{i,t}|G^t\,\pi_B) \right] \right\}$$

Parametric forms:

$P(G^t|\pi_B) \equiv P(G^t|V^{1\to t}\,O^{1\to t}\pi_A)$ is the prior taken from the result of the model of Figure 5.3;

$P(Q_C^{i,t}|G^t\,\pi_B)$ is a beta distribution $\mathcal{B}(\alpha_Q, \beta_Q)$ for $[G^t = C]$ that expresses that, for a given point of fixation proposal for the next gaze shift, $Q_C^{i,t}$ is more likely near 1, or a uniform distribution on $Q_C^{i,t}$ for $[G^t \neq C]$.

Identification:

Empirical values for free parameters α_Q and β_Q.

Questions:

$P(G^t|V^{1\to t}\,O^{1\to t}\,S^t\,\pi_B) = P(G^t|Q^t\,\pi_B)$

Figure 5.4: Bayesian Program for automatic orienting based on sensory saliency model π_B.

The first model we propose uses the knowledge from the BVM layer to determine gaze shift fixation points. More precisely, it tends to look towards locations of high entropy/uncertainty. Its likelihood will be based on the rationale conveyed by an additional variable that quantifies the uncertainty-based interest of a cell on the BVM, thus promoting entropy-based active exploration, as described on chapter 4.

The Bayesian Program for this model is presented on Fig. 5.3. The dependency of the uncertainty measure variable U_C^t — equation (4.3) — on the BVM states $(V^{1\to t}, O^{1\to t})$ are implicitly stated by definition, thus, with this model, the distribution on the point of fixation of the next desired gaze shift can be computed using the following expression

$$\begin{aligned} P(G^t|V^{1\to t}\,O^{1\to t}\,\pi_A) &= P(G^t|U^t\,\pi_A) \\ &\propto \prod_C P(U_C^t|G^t\,\pi_A) \end{aligned} \tag{5.1}$$

The second model is based on sensor models that relate sensor measurements $Z_j^{i,t}$ with $i = 1..N$ independent sensory properties of saliency ($j = 1..M_i$ total independent measurements for each saliency property), represented by the set of binary random variables $S_C^{i,t}$ (equalling 0 when the cell is non-salient and 1 when salient) corresponding to each cell C. In other words, these sensor models are generically notated as $P(Z^t|S_C^{i,t}\,V_C^t\,O_C^t\,\pi_C)$, indiscriminately of what the specific sensory saliency property $S^{i,t} = \bigwedge_C S_C^{i,t}$ might represent.

The Bayesian Program for model π_B is presented on Fig. 5.4. With this model, the distribution on the point of fixation of the next desired gaze shift can be computed using the following expression

$$\begin{aligned} &P(G^t|V^{1\to t}\,O^{1\to t}\,S^t\,\pi_B) \propto \\ &\qquad P(G^t|V^{1\to t}\,O^{1\to t}\pi_A) \prod_C \left[\prod_{i=1}^{N} P(Q_C^{i,t}|G^t\,\pi_B)\right] \end{aligned} \tag{5.2}$$

In short, this model is the product between the prior on gaze shifts due to entropy-based active exploration and each distribution on the sensory-salient cells. This expression shows that the model is attracted towards both salient cells (without necessarily looking at one in particular, as the balance between the distributions on salient cells can lead to a peak in some weighted sum of their locations) *and* locations of high uncertainty when sensory saliency is not preponderant enough (i.e. this process is called

Program

Description

Specification

Relevant variables:

O^t_C: binary value describing the occupancy of cell C at time t, $[O_C = 1]$ if cell C is occupied by an object, $[O_C = 0]$ otherwise (related variables: $O = \bigwedge_{t\in[1,t_{max}]} O^t_C$ and $O^t = \bigwedge_C O^t_C$);

V^t_C: velocity of cell C at time t, discretised into $n+1$ possible cases $\in \mathcal{V} \equiv \{v_0, \cdots, v_n\}$ (related variables: $V = \bigwedge_{t\in[1,t_{max}]} V^t_C$ and $V^t = \bigwedge_C V^t_C$);

$R^t_C \equiv f(G^{t-1}) \in [0,1]$: inhibition level for cell C modelling the Inhibition of Return behaviour (see below), ranging from no inhibition (0) to full inhibition (1) (related variables: $R^t = \bigwedge_C R^t_C$ and $R = \bigwedge_{t\in[1,t_{max}]} R^t$);

$G^t \equiv (\log_b \rho_{\max}, \theta_{\max}, \phi_{\max}) \in \mathcal{Y}$: fixation point for next gaze-shift, computed at time t, with related variable $G = \bigwedge_{t\in[1,t_{max}]} G^t$;

$S^{i,t}_C$: binary value describing the ith of N sensory saliency properties of cell C at time t, $[S^{i,t}_C = 0]$ when non-salient and $[S^{i,t}_C = 1]$ when salient (related variables: $S^i = \bigwedge_{t\in[1,t_{max}]} S^{i,t}$, $S^t = \bigwedge_{i=1}^{N} S^{i,t}$ and $S = \bigwedge_{i=1}^{N} S^i$);

$Z^{i,t}_j \in \mathcal{Z}$: sensor measurements at time t ($j = 1..M_i$ total independent measurements for each saliency property at time t, $S^{i,t}$) (related variables: $Z^i = \bigwedge_{t\in[0,t_{max}]} Z^{i,t}$ and $Z = \bigwedge_{i=1}^{N} Z^i$);

$Q^{i,t}_C = P([S^{i,t}_C = 1]|Z^{i,t}_j\, C) \in [0,1]$: probability of a perceptually salient object occupying cell C (related variables: $Q^i = \bigwedge_{t\in[1,t_{max}]} Q^{i,t}$, $Q^t = \bigwedge_{i=1}^{N} Q^{i,t}$ and $Q = \bigwedge_{i=1}^{N} Q^i$).

Decomposition:

$$P(G\, R|\pi_C) = \prod_{t=1}^{t_{max}} \left\{ P(G^t|\pi_C) \prod_C \left[P(R^t_C|G^t\, \pi_C) \right] \right\}$$

Parametric forms:

$P(R^t_C|G^t\, \pi_C)$: is a beta distribution $\mathcal{B}(\alpha_R, \beta_R)$ for $[G^t = C]$ modelling the Inhibition of Return behaviour (see main text) that expresses that, for a given point of fixation proposal for the next gaze shift, R^t_C is more likely to be 0, and a uniform distribution for $[G^t \neq C]$.

$P(G^t|\pi_C) \equiv P(G^t|V^{1\to t}\, O^{1\to t}\, S^t\, \pi_B)$ is the prior taken from the result of the model of Figure 5.4;

Identification:

Empirical values for free parameters α_R and β_R.

Questions:

$P(G^t|V^{1\to t}\, O^{1\to t}\, S^t\, G^{t-1}\pi_C) = P(G^t|R^t\, \pi_C)$

Figure 5.5: Bayesian Program for full active perception model π_C.

weighting, as opposed to *switching*, in which these behaviours would be mutually exclusive — see [Colas et al. 2010, Ferreira and Castelo-Branco 2007]).

The Bayesian Program for the third and final model π_C, which defines the full active perception hierarchy by adding an implementation the Inhibition of Return (IoR) mechanism, is presented on Fig. 5.5. With this model, the distribution on the point of fixation of the next desired gaze shift can be computed using the following expression

$$\begin{aligned} &P(G^t|V^{1\to t}\,O^{1\to t}\,S^t\,G^{t-1}\,\pi_C) \propto \\ &\quad P(G^t|V^{1\to t}\,O^{1\to t}\,S^t\,\pi_B)\prod_C \left[P(R_C^t|G^t\,\pi_C)\right] \end{aligned} \tag{5.3}$$

In conclusion, the full hierarchy, represented graphically in Fig. 5.6, is defined as the product between the prior on gaze shifts due to entropy-based active exploration and each distribution on the sensory-salient cells, while avoiding the fixation site computed on the previous time step through the IoR process, implemented by the last factor in the product. The parameters used for each distribution in this product, which define the relative importance of each level of the hierarchy and of each sensory saliency property, may be introduced directly by the programmer (like a genetic imprint) or manipulated "on the fly", which in turn allows for goal-dependent behaviour implementation (i.e. top-down influences), therefore ensuring that the framework is *adaptive.*

5.3 Implementation of an Artificial Bayesian Active Perception System

5.3.1 Visual saliency properties

Saliency properties from a generic visual cue, or, in other words, the conspicuity maps given by the BVM extended operators $Q_C^{i,t} = P([S_C^{i,t} = 1]|Z_j^{i,t}\,C) \in [0,1]$, were implemented in two steps:

1. A single-channel image with values varying between 0 and 1 is taken directly from visual cues taken from the right camera of the stereovision setup (thus simulating a dominant eye), either by directly normalising traditional dense conspicuity maps as defined by [Itti et al. 1998], or by generating a conspicuity map by forming Gaussian distributions with specific standard deviations centred on individual points of interest on the right camera image, for example in the case of sparse feature extractors such as face detection algorithms.

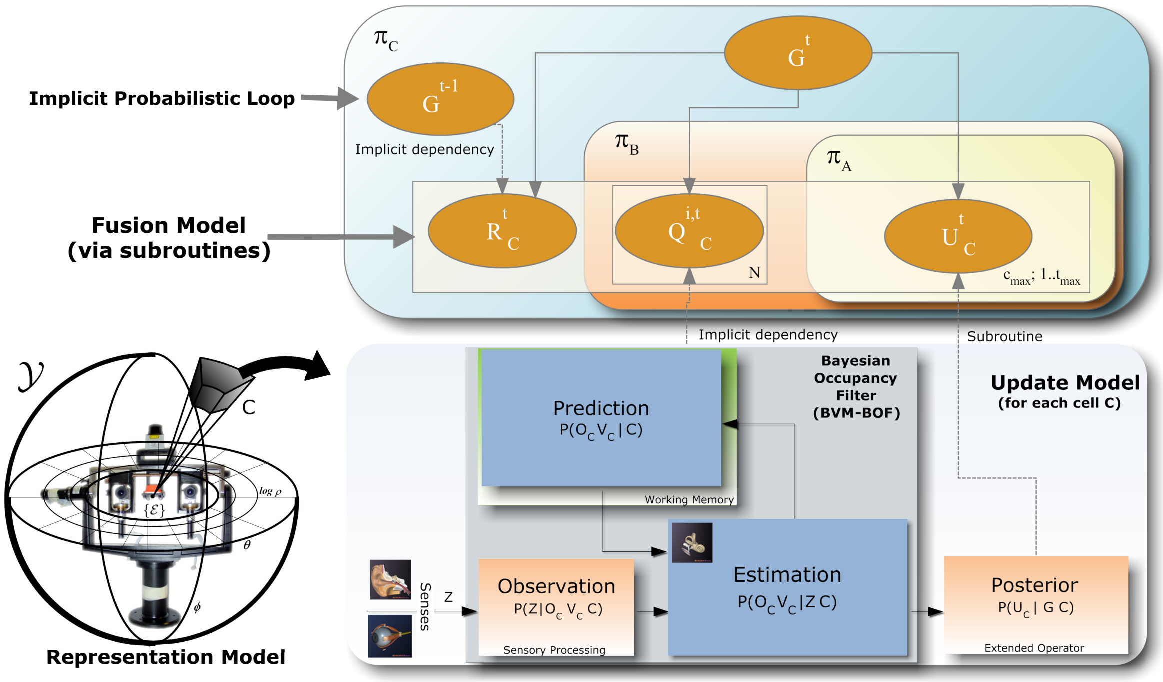


Figure 5.6: Graphical representation of the hierarchical framework for active perception. Bottom half: Update and Representation Models for the BVM-BOF framework presented in chapter 2, extended by the entropy gradient-based operator introduced in chapter 4. Upper half: Bayesian network summarising the models presented in this chapter, using the plates notation (an intuitive method of representing variables that repeat in a graphical model, so that the respective distributions appear in the joint distribution as an indexed product of the sequence of variables — for more information refer to Buntine [1994]). As can be seen, emergent behaviour results from a probabilistic fusion model implemented through a sequence of Bayesian Programming subroutines and an implicit loop that ensures the dynamic behaviour of the framework [Colas et al. 2010].

2. The saliency values from each pixel in the conspicuity map for which a disparity was estimated by the stereovision module are then projected on the log-spherical configuration through projection lines spanning the corresponding (θ, ϕ) estimates taken from equations (4.4) and (4.5) — if two or more different saliency values are projected throughout the same direction, only the highest saliency value is used. These values are thus considered as soft evidence regarding $S^{i,t}_C$, therefore yielding $Q^{i,t}_C$.

The specific properties used in this work (although any visual saliency property would have been usable by applying the two steps described above) were optical flow magnitude taken from the result of using the CUDA implementation of the “Bayesian Multi-scale Differential Optical Flow” algorithm of Simoncelli [1999] by Daniel Cabrini Hauagge (please refer to `http://www.liv.ic.unicamp.br/~hauagge/Daniel_Cabrini_Hauagge/Home_Page.html` for more information), and face detection using the Haar-like features implementation of the OpenCV library. Using these implementations, the 15 Hz performance of the stereovision unit where they are integrated reported in chapter 4 is reduced to about 6 Hz, mainly as a consequence of the slow performance of the face detection algorithm.

5.3.2 Auditory saliency properties

The auditory saliency property used in this work was directly implemented from the $P([S_C = 1]|Z\,C)$ question solved by the Bayesian model of binaural perception presented on Fig. 2.12 on page 45.

5.3.3 Inhibition of Return

The Inhibition of Return mechanism used in this work is implemented by assigning values on a log-spherical data structure corresponding to R_C^t ranging from 1 to values close to 0 depending on the distance in $\mathcal{Y}$ between G^{t-1} and each C, denoted d_{IoR}, through the following expression

$$R_C^t \equiv f(G^{t-1}) = \left(\frac{1}{2}\right)^{d_{IoR}} \tag{5.4}$$

5.3.4 Hierarchical model

The parameters distributions defined on the Bayesian Programs of Figs. 5.3, 5.4 and 5.5 were chosen for initial values in order to attain the beta distributions presented on Fig. 5.7. These preprogrammed parameters define the genetic imprint of preliminary knowledge that establishes the baseline hierarchy of the set of active perception behaviours; these parameters are changeable “on the fly” through sliders on the graphical user interface of the implementation software, thus simulating top-down influences on behaviour prioritisation (i.e. the *adaptivity* property). The influence of the relative weights imposed by these parameters will be discussed on the Results section.

The fixation point for the next time instant G^t is obtained by substituting equations (5.1) and (5.2) consecutively into (5.3), and computing $G^t = \arg\max_C P([G^t = C]|V^{1\to t}\,O^{1\to t}\,S^t\,G^{t-1}\,\pi_C)$, knowing that

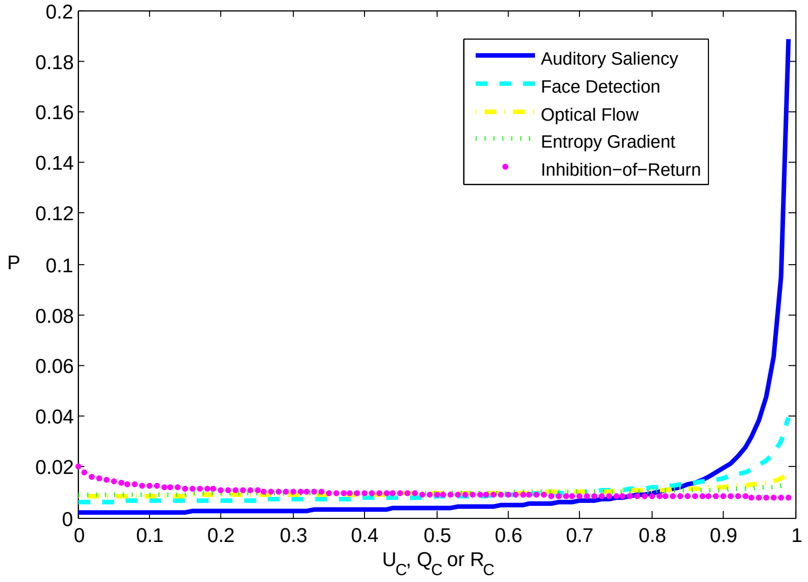


Figure 5.7: Beta distributions of the active perception hierarchy using the baseline choice for parameters. Corresponding parameters are $\alpha_U = 1$ and $\beta_U = 0.92$ for active exploration, $\alpha_Q = 1$ and $\beta_Q = 0.01$ for auditory saliency, $\alpha_Q = 1$ and $\beta_Q = 0.6$ for face detection saliency, $\alpha_Q = 1$ and $\beta_Q = 0.85$ for optical flow magnitude saliency, and $\alpha_R = 0.8$ and $\beta_R = 1$ for Inhibition of Return.

$$\begin{aligned} &P([G^t = C] | V^{1\to t}\, O^{1\to t}\, S^t\, G^{t-1}\, \pi_C) \propto \\ &\quad P(U_C^t | [G^t = C]\, \pi_A) \prod_{i=1}^{N} \left[P(Q_C^{i,t} | [G^t = C]\, \pi_B) \right] P(R_C^t | [G^t = C]\, \pi_C) \end{aligned} \tag{5.5}$$

by factoring out the effect of the uniform distributions corresponding to considering $[G^t \neq C]$.

All visual conspicuity maps were implemented using a solution similar to what was presented in chapter 4, Fig. 4.7, and sensory saliency computations as in Fig. 4.6 without resorting to aprons. These computations, all corresponding to π_B, are all performed within the auditory CUDA stream to save time. Then, in a similar way as with the saliency computations, π_A and π_C are implemented after estimating the BVM state, while the final gaze computation process is performed as described on Fig. 4.8 (b), by substituting U_C with $P([G^t = C] | V^{1\to t}\, O^{1\to t}\, S^t\, G^{t-1}\, \pi_C)$.

Finally, the full active perception system runs at about 5 Hz, for $N = 10$, $\Delta\theta = 1^{\mathrm{o}}$, $\Delta\phi = 2^{\mathrm{o}}$, mainly due to the degraded performance of the stereovision unit reported above. In any case, these ratings are still just within the parameters of satisfactory real-time performance, as defined in the concluding section of chapter 4.

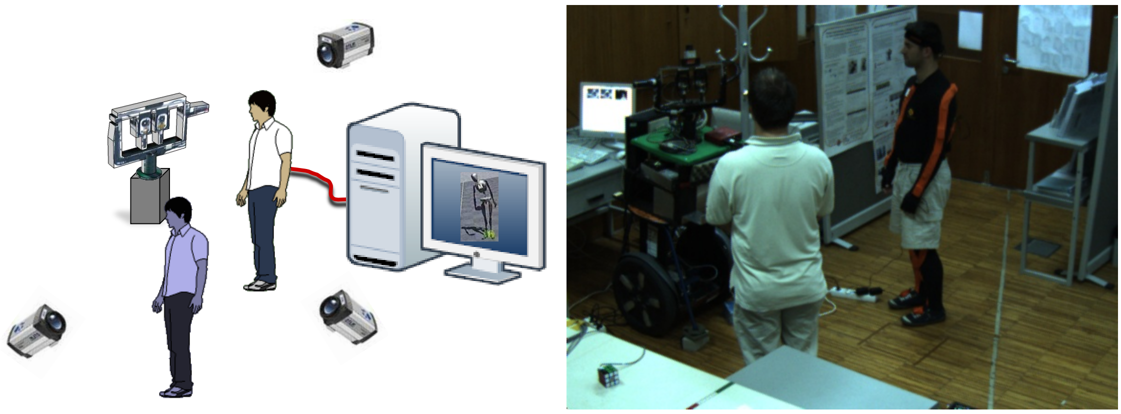


Figure 5.8: Overview of the setup used in the experimental sessions testing the Bayesian hierarchical framework for multimodal active perception. The "IMPEP 2 and interlocutors" scenario, in which one of the interlocutors is wearing body-tracking suit, is implemented using an acting script (presented on Fig. 5.9). During the experimental sessions, the signals which were recorded for analysis included data from: IMPEP 2 time-stamped video and audio logging; camera network capturing several external points of view; body-tracking poses. All signals were synchronised through common-server timestamping.

5.4 Results and Conclusions

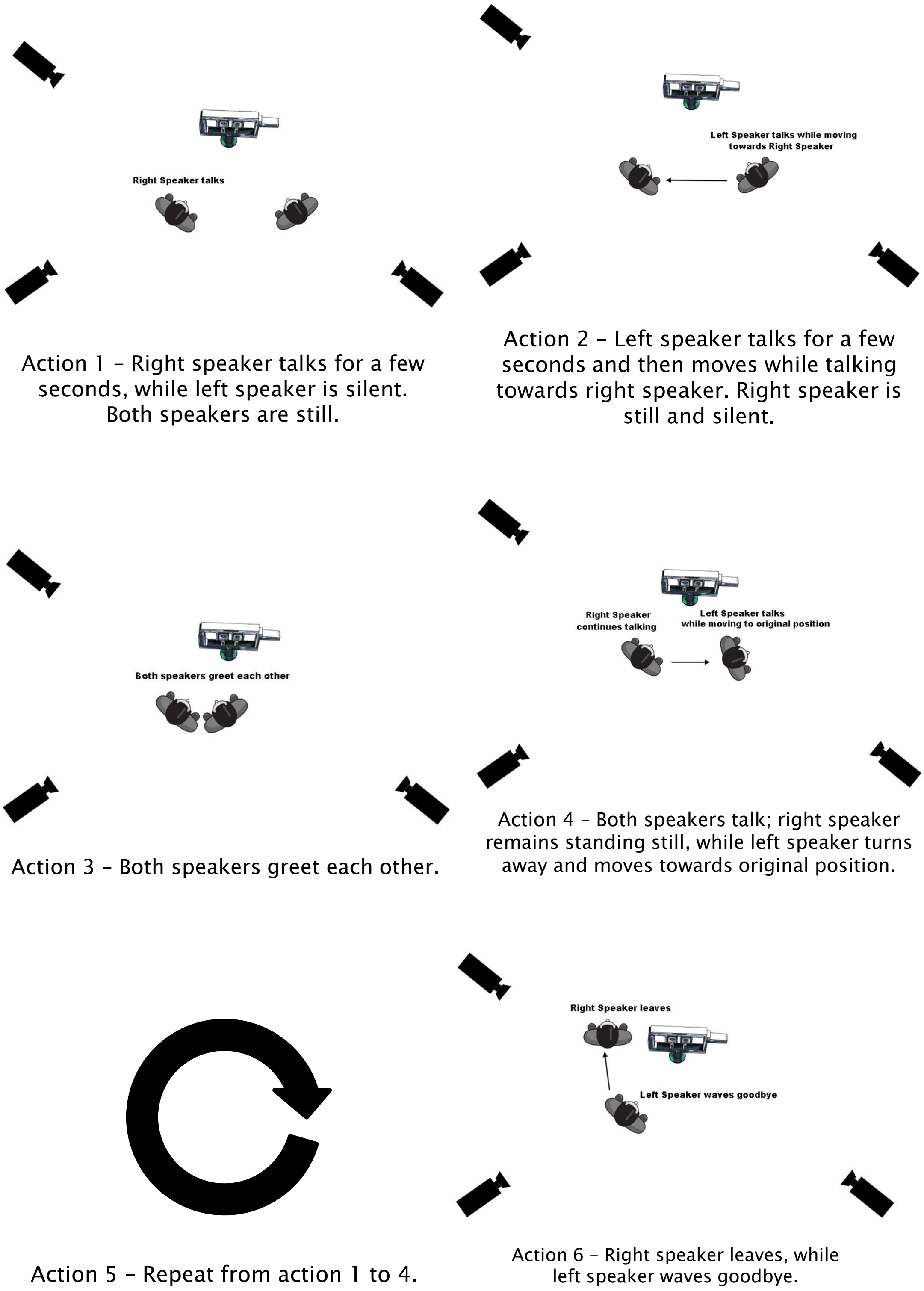
Five experimental sessions were conducted to test the performance of the hierarchical framework presented in this chapter, in particular to demonstrate its properties of emergence, scalability and adaptivity. Consequently, in the following lines, the results of each of these sessions will be discussed.

During all the experiments, three views were also filmed from external cameras — see Fig. 5.8 for an overview of the experimental setup using one of these views — and a body-tracking suit was also used by the speaker to the left from the IMPEP head's perspective, the only speaker allowed to walk from one position to another within the BVM horopter, for positioning ground-truth.

Experimental Session 1 — active perception hierarchy implementing all behaviours, using baseline priorities

In this session, a two-speaker scenario was enacted following a script (Fig. 5.9) roughly describing the activity reported in the annotated timeline of the experiment presented on Fig. 5.10.

The genetically imprinted parameters for the distributions that was used was presented on Fig. 5.7. This particular choice of parameters was made to emphasise socially-oriented, high-level behaviours as opposed to low-level behaviours and the IoR effect,


Right Speaker talks
Left Speaker talks while moving towards Right Speaker
Both speakers greet each other
Right Speaker continues talking
Left Speaker talks while moving to original position
Right Speaker leaves
Left Speaker waves goodbye

Action 1 – Right speaker talks for a few seconds, while left speaker is silent. Both speakers are still.

Action 2 – Left speaker talks for a few seconds and then moves while talking towards right speaker. Right speaker is still and silent.

Action 3 – Both speakers greet each other.

Action 4 – Both speakers talk; right speaker remains standing still, while left speaker turns away and moves towards original position.

Action 5 – Repeat from action 1 to 4.

Action 6 – Right speaker leaves, while left speaker waves goodbye.

Figure 5.9: Acting script for active perception experiments.

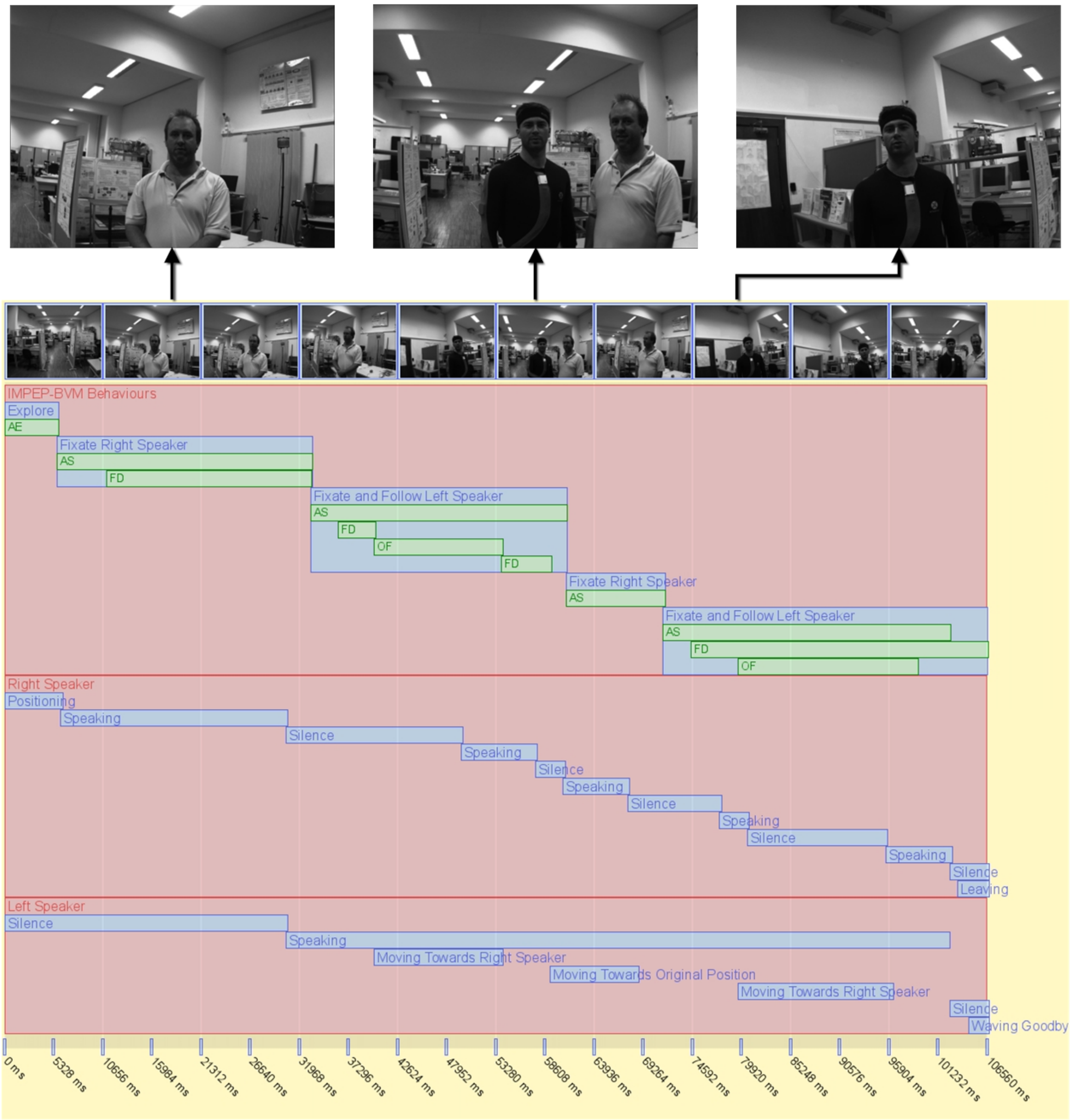


Figure 5.10: Annotated timeline for Experimental Session 1 — active perception hierarchy implementing all behaviour susing baseline priorities. The two lower annotation lanes, labelling the actions performed by the right and left speaker in the perspective of the IMPEP head, were performed by inspection of images taken by the IMPEP stereovision system, by the external cameras, by the tracking suit, and by the audio file recorded by the IMPEP binaural system. The top annotation lane, labelling the emergent behaviours of the active perception system and an interpretation of what were the most prominent underlying low-level behaviours (AE: active exploration; AS: auditory saliency; OF: optical flow magnitude saliency; FD: face detection saliency), was annotated by additionally inspecting saved logs of $P([G^t = C]|V^{1\to t}\, O^{1\to t}\, S^t\, G^{t-1}\, \pi_C)$.

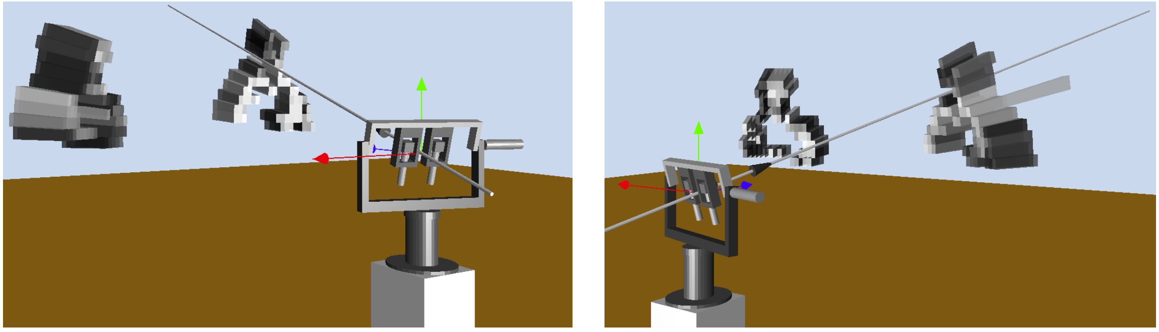


Figure 5.11: Offline rendering of a BVM representation of the two speakers scenario of Experimental Session 1. After the experiment, an instantiation of the BVM occupancy grid is rendered using a Blender-based viewer, of which two different views are presented. Notice the well-defined speaker upper torso silhouette reconstructions, which are clearly identifiable even despite the distortion elicited to visual inspection caused by the log-spherical nature of each cell. These reconstructions are better detailed as opposed to those shown on the results presented in chapter 4 due to the stabler fixation induced by the face detection saliency, allowing for accumulating more evidence on occupancy. All parameters and labelling are the same or analogous to Fig 4.15.

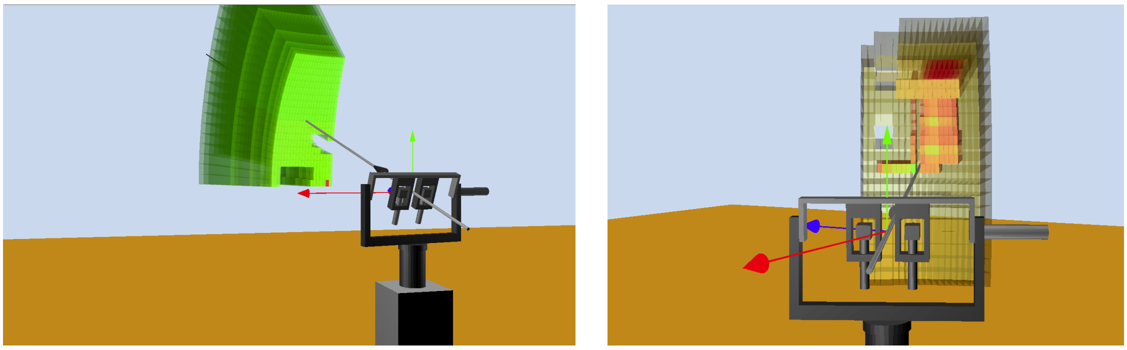


Figure 5.12: Offline rendering of example saliency maps of the two speakers scenario of Experimental Session 1. The rendering represents values for $P([G^t = C] | V^{1 \to t} \, O^{1 \to t} \, S^t \, G^{t-1} \, \pi_C)$ that were logged during the session for a specific time instant. Only a slice corresponding to all cells at 10° in azimuth and 20° in elevation around the next fixation point G^t with $P(O_C | C) > .45$ are shown, depicted using a smoothly gradated red-to-green colour-code (red corresponds to lower values, green corresponds to higher values). All other parameters and labelling are the same or analogous to Fig 5.11. On the left, a purely auditory-elicited map is shown, while on the right, a map resulting from the fusion of at least auditory and face detection conspicuity maps is shown.

which has a noticeable effect only when the former are absent. Countering the IoR effect in the presence of socially-oriented behaviours allows for an apparently more natural emergent behaviour of the system.

As can be seen in Fig. 5.10, the system successfully fixated both speakers, and even exhibited an emergent behaviour very similar to smooth pursuit while following the speaker to the left in the perspective of the IMPEP head.

This showed that the baseline priority rationale for the choice of parameters for the distributions was reasonably planned.

Offline high-definition renderings of BVM and saliency logs are presented on Figs. 5.11 and 5.12, respectively.

Experimental Session 2 — active perception hierarchy implementing all behaviours, with swapped priorities

In this session, the first part of the script of Experimental Session 1 was reenacted, but this time swapping the parameters of the distributions for auditory saliency and face detection saliency, presented on Fig. 5.7. This resulted in the system being unable to change gaze direction to the second speaker after fixating the first speaker, due to the deadlock caused by the face detection saliency keeping attention on the first speaker's face, further showcasing the importance of choosing the appropriate weights for each behaviour.

Experimental Session 3 — active perception hierarchy implementing active exploration only

In this session, the full script of Experimental Session 1 was reenacted, but this time all behaviours except entropy gradient-based active exploration were turned off by making all other distributions uniform. As expected, the behaviour described in chapter 4 emerged, namely the typical "chicken-like" saccadic movements of the IMPEP head exploring the surrounding environment, and a particular sensitivity to the entropy caused by binaural sensing and motion.

Experimental Session 4 — active perception hierarchy implementing optical flow magnitude saliency only

In this session, a single human subject (using the body-tracking suit) is tracked while walking from one position to another within the system's horopter using only optical flow magnitude saliency by making all other distributions uniform, as before.

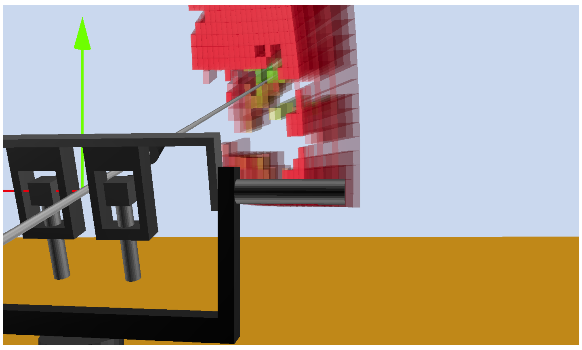


Figure 5.13: Offline rendering of an example optical flow magnitude saliency map of Experimental Session 4. All parameters and labelling are the same or analogous to Fig 5.11.

As long as the subject walked within reasonable velocity limits, the system was able to track him successfully.

A saliency map from this session, representing an example of an optical flow magnitude conspicuity map, is presented on Fig. 5.13.

Experimental Session 5 — active perception hierarchy implementing Inhibition of Return only

In this session, the IoR behaviour was tested by making all other distributions uniform, as before. In this case, a fortuitous saccadic behaviour emerged, with the system redirecting gaze to random directions at a constant rate.

In conclusion, the Bayesian hierarchical framework presented in this chapter was shown to adequately follow human-like active perception behaviours, namely by exhibiting the following desirable properties:

Emergence — High-level behaviour results from low-level interaction of simpler building blocks.

Scalability — Seamless integration of additional inputs is allowed by the Bayesian Programming formalism used to state the models of the framework.

Adaptivity — Initial "genetic imprint" of distribution parameters may be changed "on the fly" through parameter manipulation, thus allowing for the implementation of goal-dependent behaviours (i.e. top-down influences).

Chapter 6

Overall Conclusions and Future Work

6.1 Overall Conclusions

In this text we presented 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 — introducing an approach which, while not strictly neuromimetic, finds its roots in the role of the dorsal perceptual pathway of the human brain. These models build upon a common spatial configuration that is naturally fitting for the integration of readings from multiple sensors. We then presented our baseline research on human multimodal motion perception, which will serve as the foundation for future work on our framework by providing prior knowledge firmly supported by perceptual processes of the human brain. 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. Finally, we presented a real-time implementation of this system and extended the original framework by building a Bayesian hierarchical solution for multimodal active perception following human-like behaviours.

Our work on the Bayesian Volumetric Map framework has led us to the following conclusions as for its theoretical contributions:

1. The stereovision sensor model presented in this text is novel, as far as the authors know, in terms of the use of population code-like data structures to provide soft evidence in adaptive fashion (i.e. depending on an adaptive evaluation of readings taken from the environment), and provides a robust and efficient solution for

visual sensing of spatial occupancy.

2. A robust binaural sensing model that allows for the estimation of the absolute 3D positioning of sound-sources using an occupancy grid framework and based only on binaural cues is a novel approach, to the authors' knowledge.

3. A unified framework for fusion of computer vision, binaural sensing and vestibular sensing, which is, to the authors' knowledge, also novel.

4. The log-spherical configuration is designed to attain different goals in terms of spatial mapping when comparing to Euclidian solutions. In fact, to the authors' knowledge, the application of a log-spherical configuration as a solution to problems remotely similar to the ones presented in this text is also unprecedented.

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

- The results presented in chapter 4 show that the 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.

- The results presented in chapter 5 show that the full hierarchical framework exhibits the desirable properties of *emergence*, *scalability* and *adaptivity*.

- Moreover, since the human saccade-generation system promotes fixation periods (i.e. time intervals between gaze shifts) of a few hundred milliseconds on average [Carpenter 2004, Caspi et al. 2004], the overall rates of 5 to 10 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.

6.2 Future Work

6.2.1 Study of human strategies for active visuoauditory perception

The Bayesian hierarchical framework formulated earlier will now be shown to be a useful tool in proving both of our primary hypotheses for active visuoauditory perception, namely active exploration and automatic orienting using sensory saliency, as valid strategies in human behaviour regarding saccade generation through psychophysical studies. Additionally, these studies can be used to provide the data for a Bayesian learning process of switching/weighing of these strategies so as to construct an artificial active perception system that, not only generically follows, but even mimics human behaviour.

The paradigm/preliminary protocol proposed in the following lines is currently being used in a pilot experiment, which is planned to be extended to close to 20 subjects. In the following months, we hope to perform a full-fledged experiment with the final protocol based on the outcome of the pilot experiment, and consequent training and testing of the resulting active perception model using the IMPEP experimental platform.

In the future, this framework will also be used to compare the relative enacting of active perception behaviours in "healthy" subjects with subjects whose condition is believed to be directly or indirectly caused by active perception impairments, such as patients suffering from autism.

Stimuli used for this experiment consist of 3D audiovisual, stereoscopic-binaural, dynamically created on-the-fly, movies of synthetic scenes presented using an nVisor SX Head Mounted Device (HMD) by NVIS (`http://www.nvisinc.com`), with integrated eye-trackers by Arrington Research (`http://www.arringtonresearch.com/`), and also with a miniature inertial sensor, the Xsens MTi (`http://www.xsens.com/`), functioning as a head-tracker.

These scenes were generated using the NeuroVR Blender-based editor (`http://www.neurovr.org/`), and were presented continuously to each subject until each subtask wass considered to be concluded, using proprietary software developed at the ISR/FCT-UC for virtual scene visualisation and raw data logging. The subjects' tracked head-eye gaze shifts control the virtual stereoscopic-binaural point of view, and hence the progression of each stimulus movie — see Fig. 6.1. Several different scenes (i.e. with different properties and contexts) were used for each stimulus, so as to increase the amount and richness of the data used for the statistical analysis in the

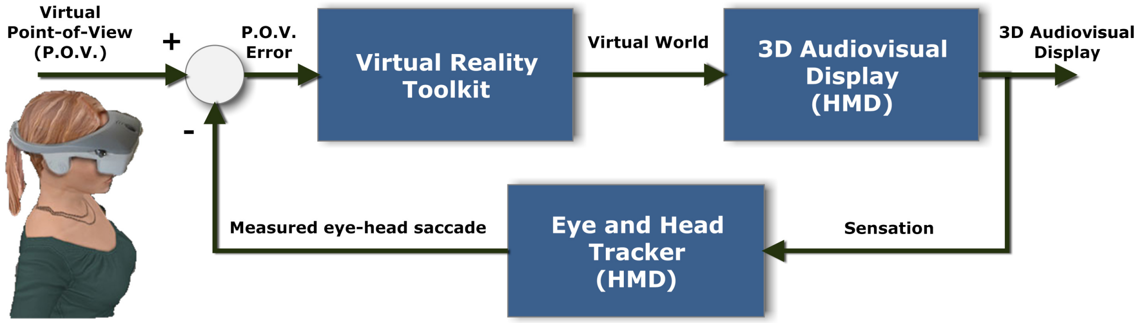


Figure 6.1: Virtual point-of-view generator setup that allows the updating of audiovisual stimuli presentation according to the monitored subjects' gaze direction.

learning process, thus increasing its statistical power.

Controlled free-viewing conditions will be enforced by proposing generic tasks to the subjects, such as "Look at the following scene; you must be able to describe it in detail when you're finished" or "Count the number of different individual objects, people and animals that you find in the following scene": these experiments thus enforced an exploratory task while avoiding implicit or personal goals that would bias the ratio between each behavioural hypothesis represented by each model.

The plausibility of these realistic presentations paired with the flexibility of the precise control of the VR world construction and editing (e.g. object and sound-source placement) in conjunction with the powerful tracking devices used for head-eye gaze shift measurements allowed for the logging of all sensory and proprioceptive data for processing by the framework and consequent comparison of the hypotheses posited by each model as described in Colas et al. [2009].

The full protocol is described on Fig. 6.2 — the relation between the protocol and the general paradigm described earlier on is depicted on Fig. 6.3.

6.2.2 Other Issues

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 dissertation.

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. [2007], but in our case using prior distributions based on multimodal perceptual integration processes, benefiting from the our baseline research

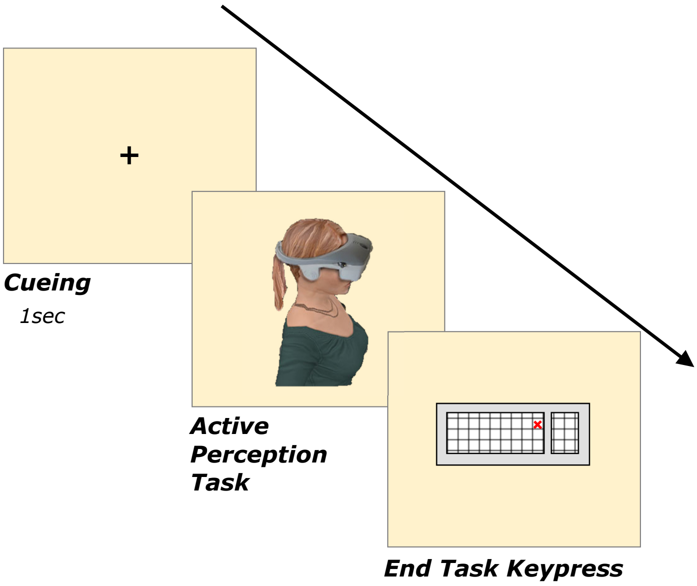


Figure 6.2: Visuoauditory-based saccade generation experimental protocol. Different virtual room stimuli are presented in N trials, with each presentation ended by the subject (by key press) whenever he/she decides that a faithful description of the surroundings can be produced, after which the next stimulus presentation takes place. During the course of each presentation, outputs from the tracking systems are recorded in a log file. The Bayesian model for saccade generation presented earlier is then fed offline with the same stimuli as the subject, the BVM filter is updated throughout time, and the gaze shift decisions made by the human subject used, together with the BVMs generated for each time instant, to perform the comparison between the decision models and hypothesis.

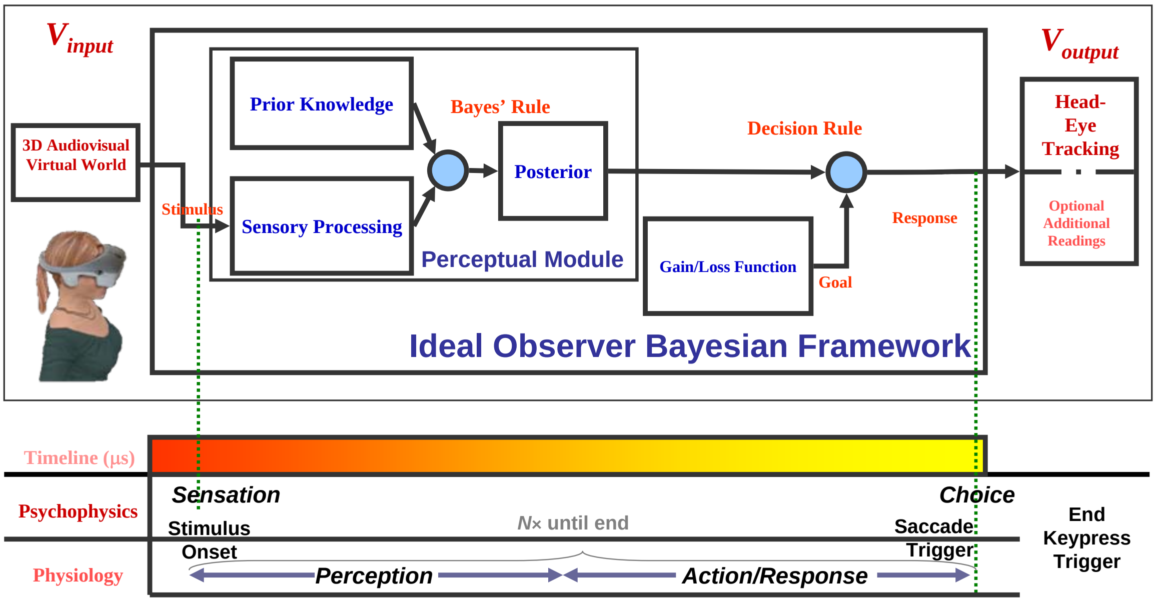


Figure 6.3: Experimental procedure schematic — adapted from the generic procedure presented on Fig. 2.2 on page 18.

on human motion perception presented on chapter 3.

On the other hand, several improvements on the CUDA implementations described in this text are still possible, in order to increase the scalability of the system and improve processing times, namely memory coalescing through pitched 2D memory operations (refer to [NVIDIA 2007] for more information), possibly the use of pinned memory on the host, and the use of multiple grids processed by parallel CUDA streams for the BVM filter in order to subdivide the BVM data structure, therefore eradicating the limit of $N = 11$ divisions in distance. Furthermore, future use of the next generation of graphics cards and CUDA Compute architectures, such as the NVIDIA Fermi [NVIDIA 2009], will make a much improved computational framework and memory subsystem available, by adding, for example, more capacity, a hierarchy with Configurable L1 and Unified L2 Caches, ECC memory support and greatly improved atomic memory operation performance.

Appendices

Appendix A

Bayesian Programming

A.1 Bayesian Program Definition

The *Bayesian Program* (BP), as first defined by Lebeltel [1999] and later consolidated by Bessière, Laugier, and Siegwart [2008], is a generic formalism for building probabilistic models and for solving decision and inference problems on these models.

This formalism was created to supersede, restate and compare numerous classical probabilistic models such as Bayesian Networks (BN), Dynamic Bayesian Networks (DBN), Bayesian Filters, Hidden Markov Models (HMM), Kalman Filters, Particle Filters, Mixture Models, or Maximum Entropy Models[1].

A Bayesian Program consists of two parts (see Fig. A.1) [Bessière et al. 2008]:

- a *description* which is the probabilistic model of the studied phenomenon or programmed behaviour;
- a *question* that specifies an inference problem to be solved using this model.

The description itself contains two subparts [Bessière et al. 2008]:

- a *specification* section that formalises the knowledge of the programmer;
- an *identification* section, in which the procedure for estimating the model's free parameters from experimental data is specified.

Some essential notation issues will be presented in the following lines.

- Random variables are represented in uppercase, such as C, and their instantiations are represented in lowercase, as in c. These instantiations are fully stated by proceeding as in the example that follows: $[C = c]$.

[1] This is detailed by Bessière et al. [2008].

$$
\text{Program}\left\{\begin{array}{l}
\text{Description}\left\{\begin{array}{l}
\text{Specification}\left\{\begin{array}{l}
\text{Relevant variables:}\\
\;X_1, X_2, \ldots, X_N\\
\text{Decomposition:}\\
\;P(X_1\, X_2\, \ldots\, X_N|\pi) =\\
\quad P(L_0|\pi)P(L_1|L_0\,\pi)\ldots P(L_K|L_{K-1}\,L_{K-2}\ldots L_0\,\pi) =\\
\quad P(L_0|\pi)P(L_1|R_1\,\pi)\ldots P(L_K|R_K\,\pi)\\
\text{Parametric forms:}\\
\;P(L_i|R_i\,\pi), \forall i \in 0\ldots K\text{: distribution of type } f_\mu(L_i),\\
\quad \text{where } \mu \text{ is a vector of parameters that may depend either on } R_i\text{, on experimental data, or both.}
\end{array}\right.\\
\text{Identification:}\\
\;\text{Method for estimating free parameters of } f_\mu(L_i).
\end{array}\right.\\
\text{Question:}\\
\;P(Search|known\,\pi)\text{, where } Search \text{ and } known \text{ (with } known \text{ being a particular instantiation of } Known\text{)}\\
\quad\text{are conjunctions of subsets of relevant variables.}
\end{array}\right.
$$

Figure A.1: Generic Bayesian Program. See main text for the definition of auxiliary variables $L_0, \ldots, L_K$ and their corresponding counterparts $R_0, \ldots, R_K$.

- Preliminary knowledge unrelated to any relevant variable[2] is expressed by the proposition denoted by π. Consequently, any proposition in a model is always conditioned by π. Therefore, in some cases π is not explicitly stated, although it is always implied.

- In general, single probability values, probability distributions and families of probability distributions are all formally denoted as *conditional probabilities*, $P(\bullet|\bullet\,\pi)$. They are distinguished from one another by the context of their arguments. For simplicity, this notation can be reduced to $P(\bullet|\bullet)$ by making the influence of hidden and latent variables implicit.

- In exceptional cases, there are only dependences on hidden or latent variables, in which case the notation reduces to $P(\bullet|\pi)$, or more simply to $P(\bullet)$.

- If the dependent variables (on the left of $|$) are not instantiated random variables or, equivalently, instantiations of random variables, this notation defines:

  - a single probability distribution if **all** variables on the right of $|$ are instantiated, or if there is only a dependence on hidden or latent variables;

[2] In other words, it is not or cannot be explicitly modelled as anything but an unknown cause of which only the effect is known, consubstantiated by hidden and latent variables (i.e. intentionally or unintentionally unaccounted for factors). Sometimes, for this reason, it is also used to formally identify a particular context for the model without explicitly describing that context.

  - otherwise, a *family* of probability distributions, one per combination of instantiations of dependent variables.

- Alternatively, if all variables are instantiated, including the dependent variables, the notation $P(\bullet|\bullet)$ (or $P(\bullet)$ if no dependency is stated) implicitly defines a single probability value (single probabilities, can exceptionally and abusively also be denoted as P_{idx}, where *idx* may be any descriptive text, for easier reading).

In the following sections, each of the constituents of a generic Bayesian Program will be explained in greater detail, based on what is presented on Bessière et al. [2008].

A.2 Description

As already defined, the description is the probabilistic model of the studied phenomenon or programmed behaviour. All the knowledge available about this phenomenon or behaviour is encoded in the *joint probability distribution* on the *relevant variables* (see Fig. A.1.

Unfortunately, this joint distribution is generally too complex to use as is. The first purpose of the description is to give an effective method of computing the joint distribution in a tractable fashion (specification). The second purpose is to specify the learning methods for identifying values of the free parameters from the observed data (identification).

A.2.1 Specification

The programmer's knowledge is specified in a sequence of three steps:

1. *Define the set of relevant variables* $\{X_1, X_2, \ldots, X_N\}$ on which the joint distribution is defined.

2. *Decompose the joint distribution* to obtain a tractable way to compute it. The only rule that must be obeyed to attain a valid probabilistic expression is that each variable must appear only once on the left side of the conditioning bar. This is formally expressed as follows. Given a partition of $\{X_1, X_2, \ldots, X_N\}$ into K subsets, we define K variables $L_0, \ldots, L_K$, each corresponding to one of these subsets. Each variable L_i is consequently obtained as the conjunction of the variables composing each subset i.

The recursive application of the *conjunction rule* of Bayesian inference [Bessière et al. 2008] leads to the exact mathematical expression on L_i presented in Fig. A.1. On the other hand, *conditional independence* hypotheses then allow for further simplifications. Such a hypothesis can be defined for variable L_i by picking a subset of variables X_j among the variables appearing in the conjunction formed by $L_{i-1} \dots L_0$ by denoting the latter by R_i and rewriting the joint distribution decomposition as shown on Fig. A.1.

3. *Define the parametric forms* that give an explicit means to compute each distribution $P(L_i|R_i\,\pi)$ appearing in the decomposition. This is achieved by associating each distribution $P(L_i|R_i\,\pi)$ with a function $f_\mu(L_i)$ — μ denotes the set of parameters that define the distribution — or a question to another Bayesian Program[3].

A.2.2 Identification

The role of the identification phase is to assign values to free parameters within the set μ, either through direct assignment or through the estimation of these parameters using *Bayesian learning* with experimental data.

A.3 Question

Given a particular description on a BP, a question is obtained by partitioning the set of relevant variables into three sets: the *searched variables* (the conjunction of which is denoted by $Search$), the *known variables* (the conjunction of which is denoted by $Known$) and the *free variables* (the conjunction of which is denoted by $Free$).

For a given value of the variable $Known$ (denoted by $known$), a question is defined as $P(Search|known\,\pi)$, as shown on Fig. A.1.

[3] This leads to two very important concepts within Bayesian programming: *subprogramming* and *hierarchies of Bayesian Programs.*

Appendix B

Bayesian Real-Time Perception Algorithms Using The Compute Unified Device Architecture (CUDA)

B.1 A Brief History of the Implementation of Perception Algorithms Using GPU Computing

GPUs have developed from fixed function architectures to programmable, multi-core architectures, leading to new applications.

A relatively popular subset of this work over the years has been vision and imaging applications. Fung and Mann [2008], present an excellent summary on this work, ranging from General Purpose GPU (GPGPU) processing, where graphics hardware is used to perform computations for tasks other than graphics, to the more recent trend of *GPU Computing*, where GPU architectures and programming tools have been developed that have created a parallel programming environment that is no longer based on the graphics processing pipeline, but still exploits the parallel architecture of the GPU — in fact, GPU Computing has transformed the GPGPU concept into the simple mapping of parellelisable algorithms onto SIMD format for the GPU, making a complete abstraction from the intricacies of graphics programming.

As a result, several full-fledged computer vision and image processing toolkits and libraries that resort to GPU technology have emerged, such as OpenVIDIA [Fung et al. 2005], GPU4Vision [GPU 2009] or GpuCV [Farrugia et al. 2006].

On the other hand, probabilistic approaches to perception have risen the stakes regarding the usefulness of GPU implementations of parallelisable algorithms. Neural network implementation is an example of this, as shown by Jang, Park, and Jung [2008], who propose a quick and efficient implementation of neural networks on both GPU and multi-core CPU, with which they developed a text detection system, achieving computational times about 15 times faster than the analogous implementation using CPU and about 4 times faster than implementation on GPU alone.

Occupancy grid-based sensor fusion algorithms, on the other hand, an example of a probabilistic approach to sensor fusion, have as of recently been a source of very interesting work on GPUs, given their obvious parallelisable trait due to the probability independence postulate between grid cells. Moreover, computational frameworks such as this are perfect candidates for GPU processing: very large data structures are processed in parallel using simple operations, yielding the perfect backdrop for SIMD-based computation. However, GPU implemetations for such algorithms are still very recent and few — examples would be the work by Reinbothe, Boubekeur, and Alexa [2009], and also Yguel, Aycard, and Laugier [2007].

Hence we believe that there is a real contribution to be made in this area, specially now, when GPU Computing has taken such a huge step forward, with the appearance of tools such as NVIDIA's CUDA architecture, which will be summarised in the following section.

B.2 The Compute Unified Device Architecture (CUDA)

We will make a brief presentation of the main features of NVIDIA's CUDA, based on the excellent summary by Hussein, Varshney, and Davis [2007]. For a detailed description, refer to [NVIDIA 2007].

B.2.1 Hardware architecture

In CUDA terminology, the GPU is called the *device* and the CPU is called the *host.* A CUDA device consists of a set of multicore processors. Each multicore processor is simply referred to as a *multiprocessor.* Cores of a multiprocessor work in a SIMD fashion. All multiprocessors have access to three common memory spaces (globally referred to as *device memory*). They are:

Constant Memory: read-only cached memory space.

Texture Memory: read-only cached memory space that is optimized for texture fetching operations.

Global Memory: read/write non-cached memory

Besides the three memory spaces that are common among all multiprocessors, each multiprocessor has an on chip *shared memory* space that is common among its cores. Furthermore, each core has an exclusive access to a read/write non-cached memory space called *local memory*.

Accessing constant and texture memory spaces is as fast as accessing registers on cache hits. Accessing shared memory is as fast as accessing registers as long as there is no bank conflict. On the other hand, accessing global and local memory spaces is much slower, typically two orders of magnitude slower than floating point multiplication and addition[1].

B.2.2 Execution model

The execution is based on *threads*. A thread can be viewed as a module, called a *kernel*, that processes a single data element of a data stream. Threads are batched in groups called *blocks*, and can only access shared memory from within their respective blocks. The group of blocks that executes a kernel constitutes one *grid*. Each thread has a three-dimensional index that is unique within its block. Each block in a grid in turn has a unique two dimensional index. Knowing its own index and the index of the block in which it resides, each thread can compute the memory address of a data element to process.

A block of threads can be executed only on a single multiprocessor. However, a single multiprocessor can execute multiple blocks simultaneously by time slicing. Threads in a block can communicate with one another via the shared memory space. They can also use it to share data fetched from global memory. There is no means of synchronization among threads in different blocks. The number of threads within a block that can execute simultaneously is limited by the number of cores in a multiprocessor. A group of threads that execute simultaneously is called a *warp*. Warps of a block are concurrently executed by time slicing.

[1] However, the new Fermi GPUs from NVIDIA will have Configurable L1 and Unified L2 caches [NVIDIA 2009].

B.2.3 Optimisation issues

There are some important considerations that need to be taken into account to obtain good performance on CUDA.

- Effect of Branching: If different threads of a warp take different paths of execution, the different paths are serialized, which reduces parallelism.
- Global Memory Read Coalescing: Global memory reads from different threads in a warp can be coalesced. To be coalesced, the threads have to access data elements in consecutive memory locations. Moreover, addresses of all data elements must follow the memory alignment guidelines. Details are in [NVIDIA 2007].
- Shared Memory Bank Conflict: Reading from shared memory is as fast as reading from registers unless a bank conflict occurs among threads. Simultaneous accesses to the same bank of shared memory are in most cases serialized.
- Writing to Global Memory: In CUDA, two or more different threads, in the same warp, can write simultaneously to the same address in global memory. The order of writing is not specified, but, one is guaranteed to succeed.

Appendix C

List of Publications

C.1 Journal Articles

Ferreira, J. F., Lobo, J., and Dias, J. Bayesian real-time perception algorithms on GPU — Real-time implementation of Bayesian models for multimodal perception using CUDA. *Journal of Real-Time Image Processing*, 26 February 2010. Published online first.

C.2 Peer-Reviewed Conference Papers

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

Ferreira, J. F. and Dias, J. A Bayesian Hierarchical Framework for Multimodal Active Perception. In *"Smarter sensors, easier processing" - Workshop, 11th International Conference on Simulation of Adaptive Behavior (SAB 2010)*, 24th August 2010.

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

Ferreira, J. F., Pinho, C., and Dias, J. Bayesian Sensor Model for Egocentric Stereovision. In *14^a Conferência Portuguesa de Reconhecimento de Padrões Coimbra (RECPAD 2008)*, October 31 2008c.

Ferreira, J. F., Pinho, C., and Dias, J. Implementation and Calibration of a Bayesian Binaural System for 3D Localisation. In *2008 IEEE International Conference on Robotics and Biomimetics (ROBIO 2008)*, Bangkok, Tailand, February, 21–26 2009a.

Ferreira, J. F., Prado, J., Lobo, J., and Dias, J. Multimodal Active Exploration Using A Bayesian Approach. In *IASTED International Conference in Robotics and Applications*, pages 319–326, Cambridge MA, USA, November 2–4 2009b.

Lobo, J., Ferreira, J. F., and Dias, J. 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, Tailand, February, 21–26 2009.

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

C.3 Technical Reports

Ferreira, J. F. and Castelo-Branco, M. 3D Structure and Motion Multimodal Perception. State-of-the-Art Report, Institute of Systems and Robotics and Institute of Biomedical Research in Light and Image, University of Coimbra, 2007. Bayesian Approach to Cognitive Systems (BACS) European Project.

Bibliography

GPU4Vision — Accelerating Computer Vision, 2009. URL `http://gpu4vision.icg.tugraz.at/`.

Alais, D., Blake, R., and Lee, S.-H. Visual features that vary together over time group together over space. *Nature Neuroscience*, 1(2):160–164, June 1998.

Alais, D. and Burr, D. The ventriloquist effect results from near optimal cross-modal integration. *Current Biology*, 14:257–262, 2004. URL `http://paloma.isr.uc.pt/bscw/bscw.cgi/d44548/Alais2004.pdf`.

Aloimonos, J., Weiss, I., and Bandyopadhyay, A. Active Vision. *International Journal of Computer Vision*, 1:333–356, 1987.

Bajcsy, R. Active perception vs passive perception. In *Third IEEE Workshop on Computer Vision*, pages 55–59, Bellair, Michigan, 1985.

Barber, M. J., Clark, J. W., and Anderson, C. H. Neural representation of probabilistic information. *Neural Computation*, 15(8):1843–1864, August 2003. ISSN 0899-7667.

Bertelson, P., Vroomen, J., de Gelder, B., and Driver, J. The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics*, 62(2):321–332, 2000.

Bessière, P., Laugier, C., and Siegwart, R., editors. *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, volume 46 of *Springer Tracts in Advanced Robotics*. Springer, 2008. ISBN: 978-3-540-79006-8.

Bohg, J., Barck-holst, C., Huebner, K., Ralph, M., Rasolzadeh, B., Song, D., and Kragic, D. Towards Grasp-oriented Visual Perception For Humanoid Robots. *International Journal of Humanoid Robotics*, 3(3):387–434, 2009.

Born, R. T. and Bradley, D. C. Structure and Function of Visual Area MT. *Annual Review of Neuroscience*, 28:157–189, July 2005.

Bouguet, J.-Y. Camera Calibration Toolbox for Matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/index.html, 2006. URL `http://www.vision.caltech.edu/bouguetj/calib_doc/index.html`.

Breazeal, C., Edsinger, A., Fitzpatrick, P., and Scassellati, B. Active Vision for Sociable Robots. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 31(5):443–453, September 2001.

Breazeal, C. and Scassellati, B. A Context-Dependent Attention System for a Social Robot. In *Sixteenth International Joint Conference on Artificial Intelligence table of contents*, pages 1146 – 1153, 1999.

Buntine, W. L. Operations for Learning with Graphical Models. *Journal of Artificial Intelligence Research (AI Access Foundation)*, 2:159–225, December 1994. ISSN 11076-9757.

Burr, D. and Alais, D. Combining visual and auditory information. *Prog Brain Res*, 155:243–258, 2006.

Burr, D. C., Morrone, M. C., and Ross, J. Selective suppression of the magnocellular visual pathway during saccadic eye movements. *Nature*, 371:511–513, 6 October 1994. Letters to nature.

Calamia, P. T. Three-dimensional localization of a close-range acoustic source using binaural cues. Master's thesis, Faculty of the Graduate School of The University of Texas at Austin, 1998.

Carmi, R. and Itti, L. Causal saliency effects during natural vision. In *ACM Eye Tracking Research and Applications*, pages 1–9, 2006.

Carpenter, R. H. S. The neural control of looking. *Current Biology*, 10:291–293, 2000. Primer.

Carpenter, R. H. S. The saccadic system: a neurological microcosm. *Advances in Clinical Neuroscience and Rehabilitation*, 4:6–8, 2004. Review Article.

Carrasco, M., Talgar, C., and Cameron, E. L. Characterizing visual performance fields: Effects of transient covert attention, spatial frequency, eccentricity, task and set size. *Spatial Vision*, 15:61–75, 2001.

CASPI, A., BEUTTER, B. R., AND ECKSTEIN, M. P. The time course of visual information accrual guiding eye movement decisions. *Proceedings of the National Academy of Sciences U.S.A.*, 101(35):13086–13090, 31 August 2004.

CASTELO-BRANCO, M. Neural correlates of visual hallucinatory phenomena: The role of attention. *Behavioral and Brain Sciences*, 28(6):760–761, 2005.

CASTELO-BRANCO, M., FORMISANO, E., BACKES, W., ZANELLA, F., NEUENSCHWANDER, S., SINGER, W., AND GOEBEL, R. W. Activity patterns in human motion-sensitive areas depend on the interpretation of global motion. *Proceedings of the National Academy of Sciences*, 99:13914–13919, October 15 2002.

CASTELO-BRANCO, M., GOEBEL, R., NEUENSCHWANDER, S., AND SINGER, W. Neural Synchrony Correlates With Surface Segregation Rules. *Nature*, 405:685–689, 2000.

CASTELO-BRANCO, M., MENDES, M., SILVA, M. F., JANUÁRIO, C., MACHADO, E., PINTO, A., FIGUEIREDO, P., AND FREIRE, A. Specific retinotopically based magnocellular impairment in a patient with medial visual dorsal stream damage. *Neuropsychologia*, 44:238–253, 2006.

CASTELO-BRANCO, M., MENDES, M., SEBASTIÃO, A. R., REIS, A., SOARES, M., SARAIVA, J., BERNARDES, R., FLORES, R., PÉREZ-JURADO, L., AND SILVA, E. Visual phenotype in Williams-Beuren syndrome challenges magnocellular theories explaining human neurodevelopmental visual cortical disorders. *Journal of Clinical Investigation*, 117(12):3720–3729, 2007.

COLAS, F., DIARD, J., AND BESSIÈRE, P. Common Bayesian Models For Common Cognitive Issues. *Acta Biotheoretica*, 58(2–3):191–216, July, 24th 2010.

COLAS, F., FLACHER, F., TANNER, T., BESSIÈRE, P., AND GIRARD, B. Bayesian models of eye movement selection with retinotopic maps. *Biological Cybernetics*, 100: 203–214, 2009.

COUÉ, C., PRADALIER, C., LAUGIER, C., FRAICHARD, T., AND BESSIÈRE, P. Bayesian occupancy filtering for multitarget tracking: an automotive application. *Int. Journal of Robotics Research*, 25(1):19–30, 2006.

CRAWFORD, J. D., CEYLAN, M. Z., KLIER, E. M., AND GUITTON, D. Three-Dimensional Eye-Head Coordination During Gaze Saccades in the Primate. *Journal of Neurophysiology*, 81:1760–1782, 1999. The American Physiological Society.

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

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

Dankers, A., Barnes, N., and Zelinsky, A. Active Vision for Road Scene Awareness. In *IEEE Intelligent Vehicles Symposium (IVS05)*, Los Vegas, USA, June 2005.

de Boer, E. Synthetic whole-nerve action potentials for the cat. *Acoustical Society of America Journal*, 58:1030–1045, November 1975.

de Croon, G. C. H. E., Sprinkhuizen-Kuyper, I. G., and Postma, E. O. Comparing Active Vision Models. *Image and Vision Computing*, 27(4):374–384, 2009.

Denève, S., Latham, P. E., and Pouget, A. Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience*, 2(8):740–745, 1999.

Denève, S. and Pouget, A. Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology — Paris*, 98:249–258, 2004.

Dyde, R. T. and Milner, A. D. Two illusions of perceived orientation: one fools all of the people some of the time; the other fools all of the people all of the time. *Experimental Brain Research*, 144:518–527, 2002.

Elazary, L. and Itti, L. Interesting objects are visually salient. *Journal of Vision*, 8(3):1–15, 2008.

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

Elfes, A. Multi-Source Spatial Data Fusion Using Bayesian Reasoning. In Abidi, M. A. and Gonzalez, R. C., editors, *Data Fusion in Robotics and Machine Intelligence*. Academic Press, 1992.

Ernst, M. O. and Bülthoff, H. H. Merging the senses into a robust percept. *TRENDS in cognitive Sciences*, 8(4):162–169, April 2004.

Faller, C. and Merimaa, J. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *Journal of the Acoustical Society of America*, 116(5):3075–3089, November 2004a.

Farivar, R. Dorsal-ventral integration in object recognition. *Brain Research Reviews*, 2:144–153, October 2009.

Farrugia, J.-P., Horain, P., Guehenneux, E., and Alusse, Y. GpuCV: A framework for image processing acceleration with graphics processors. In *2006 IEEE International Conference on Multimedia and Expo*, pages 585–588, July 9–12 2006.

Fung, J. and Mann, S. Using Graphics Devices in Reverse: GPU-based Image Processing and Computer Vision. In *IEEE Int'l Conf. on Multimedia and Expo*, Hannover, Germany, June 23–26 2008.

Fung, J., Mann, S., and Aimone, C. OpenVIDIA: Parallel GPU Computer Vision. In *ACM Multimedia 2005*, pages 849–852, Singapore, November 6–11 2005.

Galati, G., Lobel, E., Vallar, G., Berthoz, A., Pizzamiglio, L., and Bihan, D. L. The neural basis of egocentric and allocentric coding of space in humans: a functional magnetic resonance study. *Experimental Brain Research*, 133:156–164, April 2000.

Gallup, D. CUDA Stereo. `http://www.cs.unc.edu/~gallup/stereo-demo`, 2009.

Geisler, W. S. Ideal-observer theory in psychophysics and physiology. *Physica Scripta*, 39:153–160, 1989a.

Geisler, W. S. Sequential ideal-observer analysis of visual discrimination. *Psychological Review*, 96:267–314, 1989b.

Geisler, W. S. Ideal Observer analysis. In Chalupa, L. and Werner, J., editors, *The Visual Neurosciences*, pages 825–837. MIT press, 2003.

Geisler, W. S. and Kersten, D. Illusions, perception and Bayes. *Nature Neuroscience*, 5(6):508–510, June 2002.

Goodale, M. A. and Milner, A. D. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.

Gordon, J., Ghilardi, M. F., and Ghez, C. Accuracy of planar reaching movements. I. Independence of direction and extent variability. *Experimental Brain Research*, 99(1):97–111, 1994.

Henkel, R. D. A Simple and Fast Neural Network Approach to Stereovision. In Jordan, M., Kearns, M., and Solla, S., editors, *Proceedings of the Conference on Neural Information Processing Systems — NIPS'97*, pages 808–814, Denver, 1998. MIT Press, Cambridge.

Henkel, R. D. Synchronization, Coherence-Detection and Three-Dimensional Vision. Technical report, Institute for Theoretical Physics, 2000.

Hussein, M., Varshney, A., and Davis, L. On Implementing Graph Cuts on CUDA. In *First Workshop on General Purpose Processing on Graphics Processing Units*, Boston, MA, October 2007.

Immerseel, L. V. and Peeters, S. Digital implementation of linear gammatone filters: Comparison of design methods. *Acoustics Research Letters Online*, 4(3): 59–64, July 2003. Published online by the Acoustical Society of America.

Itti, L. and Baldi, P. Bayesian surprise attracts human attention. *Vision Research*, 49:1295–1306, 2009.

Itti, L., Koch, C., and Niebur, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.

Jacobs, R. A. What determines visual cue reliability? *TRENDS in Cognitive Sciences*, 6(8):345–350, 2002. Review.

Jain, A., Sally, S. L., and Papathomas, T. V. Audiovisual short-term influences and aftereffects in motion: Examination across three sets of directional pairings. *Journal of Vision*, 8(15):1–13, 2008.

Jang, H., Park, A., and Jung, K. Neural Network Implementation Using CUDA and OpenMP. In *Proceedings of the 2008 Digital Image Computing: Techniques and Applications*, pages 155–161. IEEE Computer Society Washington, DC, USA, 2008.

Jaynes, E. T. *Probability Theory: the Logic of Science*. Cambridge University Press, 2003.

Jazwinsky, A. H. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970. ISBN 0-12381-5509.

Kapralos, B., Jenkin, M. R. M., and Milios, E. Auditory Perception and Spatial (3D) Auditory Systems'. Technical Report CS-2003-07, York University, July 2003.

King, A. J., Schnupp, J. W., and Doubell, T. P. The shape of ears to come: dynamic coding of auditory space. *TRENDS in Cognitive Sciences*, 5(6):261–270, 2001.

Knill, D. C. and Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, December 2004.

Knudsen, E. and Brainard, M. Creating a unified representation of visual and auditory space in the brain. *Annual Review of Neuroscience*, 18:19–43, 1995. W. Maxwell Cowan ed.

Koene, A., Morén, J., Trifa, V., and Cheng, G. Gaze shift reflex in a humanoid active vision system. In *5th International Conference on Computer Vision Systems (ICVS 2007)*, Bielefeld University, Germany, 2007. Applied Computer Science Group. ISBN 978-3-00-020933-8.

Kopp, L. and Gärdenfors, P. Attention as a minimal criterion of intentionality in robots. *Cognitive Science Quarterly*, 2:302–319, 2002.

Kozak, L. R. and Castelo-Branco, M. Peripheral influences on motion integration in foveal vision are modulated by central local ambiguity and center-surround congruence. *Investigative Ophthalmology & Visual Science*, October 24 2008. Published online ahead of print.

Laurens, J. and Droulez, J. Bayesian modeling of inertial self-motion perception. Section 3.2 of Bayesian IBA Project Workpackage 2 deliverable D15, 2005.

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

Lebeltel, O. *Programmation Bayésienne des Robots*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, September 1999.

Liu, T., Heeger, D. J., and Carrasco, M. Neural correlates of visual vertical meridian asymmetry. *Journal of Vision*, 6:1294–1306, 2006.

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

Lobo, J. and Dias, J. Relative Pose Calibration Between Visual and Inertial Sensors. *International Journal of Robotics Research, Special Issue 2nd Workshop on Integration of Vision and Inertial Sensors*, 26(6):561–575, June 2007.

López-Moliner, J. and Soto-Faraco, S. Vision affects how fast we hear sounds move. *Journal of Vision*, 7(12):1–7, 2007.

Lu, Y.-C., Christensen, H., and Cooke, M. Active binaural distance estimation for dynamic sources. In *Interspeech 2007*, Antwerp, Belgium, 2007.

McIntyre, J., Stratta, F., and Lacquaniti, F. Short-Term Memory for Reaching to Visual Targets: Psychophysical Evidence for Body-Centered Reference Frames. *Journal of Neuroscience*, 18(20):8423–8435, October 15 1998.

Meyer, G. F. and Wuerger, S. M. Cross-modal integration of auditory and visual motion signals. *NeuroImage*, 12(11):2557–2560, August 2001.

Meyer, G. F., Wuerger, S. M., Rohrbein, F., and Zetzsche, C. Low-level integration of auditory and visual motion signals requires spatial co-localisation. *Exp Brain Res*, 166:538–547, 2005.

Mishkin, M., Ungerleider, L. G., and Macko, K. A. Object vision and spatial vision: two cortical pathways. *Trends in Neuroscience*, 6:414–417, October 1983.

Montemerlo, M., Roy, N., Thrun, S., Haehnel, D., Stachniss, C., and Glover, J. Carmen, the carnegie mellon robot navigation toolkit. Downloaded from http://carmen.sourceforge.net/home.html, 2007. Open-source collection of software for mobile robot control.

Moravec, H. and Elfes, A. High resolution maps from wide angle sonar. In *IEEE International Conference on Robotics and Automation*, 1985.

Moravec, H. P. Sensor fusion in certainty grids for mobile robots. *AI Magazine*, 9 (2):61–74, 1988.

Mozolic, J. L., Hugenschmidt, C. E., and Peiffer, A. M. Modality-specific selective attention attenuates multisensory integration. *Experimental Brain Research*, 184:39–52, 2008.

Murphy, K. J., Carey, D. P., and Goodale, M. A. The Perception of Spatial Relations in a Patient with Visual Form Agnosia. *Cognitive Neuropsysholology*, 15 (6/7/8):705–722, 1998.

Niebur, E., Itti, L., and Koch, C. Modeling the "where´´ visual pathway. In Sejnowski, T. J., editor, *2nd Joint Symposium on Neural Computation, Caltech-UCSD*, volume 5, pages 26–35, La Jolla, 1995. Institute for Neural Computation.

NVIDIA. CUDA Programming Guide ver 1.2, 2007.

NVIDIA. NVIDIA's Next Generation CUDA^{TM} Compute Architecture: Fermi^{TM}. Whitepaper, NVIDIA, 2009. Published online: `http://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf`.

Oppenheim, A. V. and Schafer, R. *Discrete-Time Signal Processing*. 1989.

Ouerhani, N., von Wartburg, R., Hugli, H., and Muri, R. Empirical Validation of the Saliency-based Model of Visual Attention. In *Electronic Letters on Computer Vision and Image Analysis 3(1):13-24,*, 2004.

Pagac, D., Nebot, E., and Durrant-Whyte, H. An evidential approach to map-building for autonomous vehicles. *IEEE Trans. on Robotics and Automation*, 14(4):623–629, 1998.

Parkhurst, D., Law, K., and Niebur, E. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107–123, 2002.

Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice. SVOS Final Report: The auditory filterbank. Technical Report 2341, MRC Applied Psychology Unit, Cambridge., 1988.

Patterson, R. D., Allerhand, M. H., and Giguère, C. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *J. AcoustS. Soc. Am*, pages 1890–1894, 1995.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc. (Elsevier), revised second printing edition, 1988.

Pizlo, Z. Perception viewed as an inverse problem. *Vision Research*, 41(24):3145–3161, November 2001.

Poggio, T. Vision by man and machine. *Scientific American*, 250(4):106–116, April 1984.

Pouget, A., Dayan, P., and Zemel, R. Information processing with population codes. *Nature Reviews Neuroscience*, 1:125–132, 2000. Review.

Previc, F. H. Functional specialization in the lower and upper visual fields in humans: Its ecological origins and neurophysiological implications. *Behavioral and Brain Sciences*, 13:519–575, 1990.

Pulkki, V. Creating generic soundscapes in multichannel loudspeaker systems using vector base amplitude panning in Csound synthesis software. *Organised Sound*, 3 (2):129–134, 1998.

Rao, R. P. N. Bayesian Computation in Recurrent Neural Circuits. *Neural Computation*, 16(1):1–38, 2004. ISSN 0899-7667.

Rao, R. P. N. Bayesian inference and attentional modulation in the visual cortex. *NeuroReport — Cognitive Neuroscience and Neurophysiology*, 16(16):1843–1848, November 2005. ISSN 0899-7667.

Reinbothe, C., Boubekeur, T., and Alexa, M. Hybrid Ambient Occlusion. In *Proceedings of the Eurographics Symposium on Rendering*, 2009.

Rocha, R. *Building Volumetric Maps with Cooperative Mobile Robots and Useful Information Sharing: a Distributed Control Approach based on Entropy.* PhD thesis, Faculty of Engineering of the University of Porto, 2005.

Rocha, R., Dias, J., and Carvalho, A. Cooperative Multi-Robot Systems: a study of Vision-based 3-D Mapping using Information Theory. *Robotics and Autonomous Systems*, 53(3–4):282–311, December 2005a.

Rocha, R., Dias, J., and Carvalho, A. Exploring information theory for vision-based volumetric mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'2005)*, pages 2409–2414, Edmonton, Canada, August 2005b.

Rommelsea, N. N., der Stigchelc, S. V., and Sergeant, J. A. A review on eye movement studies in childhood and adolescent psychiatry. Article in press; corrected proof, 2008.

Rubin, N., Nakayama, K., and R, R. S. Enhanced perception of illusory contours in the lower versus upper visual hemifields. *Science*, 27:651–653, 1996.

Schmidt, K. E., Castelo-Branco, M., Goebel, R., Payne, B. R., Lomber, S. G., and Galuske, R. A. Pattern motion selectivity in population responses of area 18. *European Journal of Neuroscience*, 24:2363–74, 2006.

Schrater, P. and Kersten, D. Vision, Psychophysics and Bayes. In Rao, R. P. N., Olshausen, B. A., and Lewicki, M. S., editors, *Probabilistic Models of the Brain*, chapter 2, pages 37–60. MIT Press, Cambridge, MA, 2001.

Shibata, T., Vijayakumar, S., Conradt, J., and Schaal, S. Biomimetic Oculomotor Control. *Adaptive Behaviour - Special Issue on Biologically Inspired and Biomimetic System*, 9(3–4):189–208, 2001.

Shic, F. and Scassellati, B. A Behavioral Analysis of Computational Models of Visual Attention. *International Journal of Computer Vision*, 73(2):159–177, 2007.

Shih, S.-W., Hung, Y.-P., and Lin, W.-S. Calibration of an Active Binocular Head. *IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans*, 28(4):426–442, July 1998.

Shinn-Cunningham, B. G., Santarelli, S., and Kopco, N. Tori of confusion: Binaural localization cues for sources within reach of a listener. *Journal of the Acoustical Society of America*, 107(3):1627–1636, March 2000.

Silva, M. F., Maia-Lopes, S., Mateus, C., Guerreiro, M., Sampaio, J., Faria, P., and Castelo-Branco, M. Retinal and cortical patterns of spatial anisotropy in contrast sensitivity tasks. *Vision Research*, 48(1):127–35, 2008.

Silver, M. A. and Kastner, S. Topographic maps in human frontal and parietal cortex. *Trends in cognitive sciences*, 13:488–495, November 2009.

Simoncelli, E. P. *Bayesian Multi-Scale Differential Optical Flow*, volume 2 of *Handbook of Computer Vision and Applications*, chapter 14. Academic Press, 1999.

Soto-Faraco, S., Kingstone, A., and Spence, C. Multisensory contributions to the perception of motion. *Neuropsychologia*, 41:1847–1862, 2003.

Soto-Faraco, S., Spence, C., Lloyd, D., and Kingstone, A. Moving Multisensory Research Along: Motion Perception Across Sensory Modalities. *Current Directions in Psychological Science*, 13(1):29–32, 2004.

Sparks, D. L. Conceptual issues related to the role of the superior colliculus in the control of gaze. *Current Opinion in Neurobiology*, 9:698–707, 1999.

Spence, C. and Driver, J. Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics*, 59:1–22, 1997.

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

Stachniss, C. and Burgard, W. Mapping and exploration with mobile robots using coverage maps. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 467–472, 2003.

Tatler, B. W., Baddeley, R. J., and Gilchrist, I. D. Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5):643–659, 2005a.

Tay, C., Mekhnacha, K., Chen, C., Yguel, M., and Laugier, C. An efficient formulation of the bayesian occupation filter for target tracking in dynamic environments. *International Journal of Autonomous Vehicles*, 2007.

Treue, S., Hol, K., and Rauber, H.-J. Seeing multiple directions of motion — physiology and psychophysics. *Nature Neuroscience*, 3(3):270–276, March 2000.

Tsotsos, J. and Shubina, K. Attention and Visual Search : Active Robotic Vision Systems that Search. In *The 5th International Conference on Computer Vision Systems*, Bielefeld, March 21 - 24 2007.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., and Nuflo, F. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78:507–545, 1995.

Turano, K. A., Geruschat, D. R., and Baker, F. H. Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research*, 43:333–346, 2003.

Ungerleider, L. G. and Mishkin, M. Two cortical visual systems. In Ingle, D. J., Goodale, M. A., and Mansfield, R. J. W., editors, *Analysis of visual behaviour*. MIT Press, Cambridge, M A, 1982.

Viemeister, N. F. and Wakefield, G. H. Temporal integration and multiple looks. *The Journal of the Acoustical Society of America*, 90(2):858–865, August 1991.

Vroomen, J. and de Gelder, B. Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance*, 26:1583–1590, 2000.

Wallace, M. T., Meredith, M. A., and Stein, B. E. Multisensory Integration in the Superior Colliculus of the Alert Cat. *Journal of Neurophysiology*, 80(2):1006–1001, August 1998. The American Physiological Society, Rapid Communication.

Watson, T. L. and Krekelberg, B. The Relationship between Saccadic Suppresion and Perceptual Stability. *Current Biology*, 19(12):1040–1043, 23 June 2009.

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604, 2002.

Welch, G. and Bishop, G. An Introduction to the Kalman Filter. Technical Report TR 95-041, Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3175, July 2006.

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

Yuille, A. L. and Bülthoff, H. H. Bayesian decision theory and psychophysics. In Knill, D. and Richards, W., editors, *Perception as Bayesian Inference*, pages 123–161. Cambridge University Press, Cambridge, December 1996.

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

Zemel, R. S., Dayan, P., and Pouget, A. Probabilistic Interpretation of Population Codes. *Advances in Neural Information Processing Systems*, 9:676–683, 1997.

Zhou, F., Wong, V., and Sekuler, R. Multi-sensory integration of spatio-temporal segmentation cues: one plus one does not always equal two. *Experimental Brain Research*, 180(4):641–654, 2007.

Zurek, P. M. The precedence effect. In Yost, W. and G.Gourevitch, editors, *Directional Hearing*. Sprigler-Verlag, 1987.

Errata

Page 3, Paragraph 2, from Line 3 onwards:

"Aided by developments… …of perceptual variables [Knill and Pouget 2004]." should read:

"Recent advances both in statistics and artificial intelligence have spurred researchers to begin 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. This observation, along with the behavioural and computational work on which it is based, has fundamental implications for neuroscience, particularly in how we conceive of neural computations and the nature of neural representations of perceptual variables» [Knill and Pouget 2004]."

Page 3, Paragraph 3:

This paragraph should read:

"Human perception is clearly not optimal; humans only achieve the level of performance afforded by the uncertainty in the physical stimulus [Knill and Pouget 2004; Kozak and Castelo-Branco 2008]. «Absolute efficiencies (a measure of performance relative to a Bayesian optimal observer) for performing high-level perceptual tasks are generally low and vary widely across tasks» [Knill and Pouget 2004] – see [Silva et al. 2008] for such an example. «In some cases, this inefficiency is entirely due to uncertainty in the coding of sensory primitives that serve as inputs to perceptual computations; in others, it is due to a combination of sensory, perceptual and cognitive factors. The real test of the Bayesian coding hypothesis is in whether the neural computations that result in perceptual judgements or motor behaviour take into account the uncertainty in the information available at each stage of processing» [Knill and Pouget 2004]. Psychophysical work in several areas suggests that this is the case [Knill and Pouget 2004], being particularly evident in clinical models [Castelo-Branco, Mendes, Sebastião, Reis, Soares, Saraiva, Bernardes, Flores, Pérez-Jurado, and Silva 2007]."

Page 6, Paragraphs 1 and 2:

These paragraphs should be enclosed in quotes (**«»**).

Page 6, Paragraph 4, continued to page 7:

This paragraph should be enclosed in quotes (**«»**).

Page 7, Paragraph 3, from Line 2:

The period reading "allocentric spatial… …the visual control of action [Murphy et al. 1998]" should read **"«allocentric spatial… …the visual control of action» [Murphy et al. 1998]."**

Page 7, Paragraph 4, from Line 3:

The period reading "An important fact about results… …[Cutting and Vishton 1995]" should read **"«An important fact about results… …» [Cutting and Vishton 1995]."**

Page 8, Paragraph 1, from Line 9:

The period reading "Distance or depth errors… … at different positions [Cutting and Vishton 1995]" should read **"Errors in estimating depth most probably occur in distant portions of the visual field, since in this case depth cues are below a detectable threshold and thus unable to support the perception of depth between objects at different positions [Cutting and Vishton 1995]."**

Page 95, Paragraph 2, from Line 2:

The period reading "In a subject's point of view,… (i.e., *covert attention*)." should read **"In a subject's point of view, gaze fixation may be switched to the point being attended to (i.e., *overt attention*) or, alternatively, attentional processing may also be switched without involving any fixation shift or motor action (i.e., *covert attention*)."**

Page 95, Paragraph 3, from Line 5 (continued in Page 96):

The period reading "There is now… …under debate [Parkhurst et al. 2002]." should be enclosed in quotes (**«»**).

Page 96, Paragraph 6:

The period reading "Itti et al.'s model is a… … reproducibility of results [Shic and Scassellati 2007]." should read **"Itti et al.'s model «is a… … reproducibility of results» [Shic and Scassellati 2007]."**.

Acronyms and Abbreviations

2D	**two-dimensions**/two-dimensional
3D	**three-dimensions**/three-dimensional
AIM	**a**uditory image model
AKG	Akustische und Kino-Geräte Gesellschaft m.b.H. (company)
BACS	Bayesian Approach to Cognitive Systems (European Project)
BMM	basilar membrane motion
BOF	Bayesian Occupancy Filter
BN	Bayesian Network
BP	Bayesian Program
BVM	Bayesian Volumetric Map
CPU	central processing unit
CUDA	Compute Unified Device Architecture
ECC	Error-correcting code
ERB	equivalent rectangular bandwidths
FEF	frontal eye fields
GPGPU	General Purpose GPU processing
fMRI	functional magnetic resonance imaging
EC	European Commission
FCT-UC	Faculty of Sciences and Technology of the University of Coimbra
GLM	generalised linear model
GPU	graphics processing unit
IC	interaural coherence
ICc	central nucleus of the inferior colliculus
ICx	external nucleus of the inferior colliculus
IEEE	Institute of Electrical and Electronics Engineers
ILD	interaural level difference
IMPEP	Integrated Multimodal Perception Experimental Platform
IMU	inertial measuring unit
IoR	Inhibition of Return
ISR	Institute of Systems and Robotics (University of Coimbra)
ITD	interaural time difference
MAP	maximum *a posteriori*
NAP	neural activity pattern
OpenCV	Open Source Computer Vision Library
POP	Perception on Purpose (European Project)
SIMD	single instruction multiple data
SOC	superior olivary complex
SC	superior colliculus