TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Space-Variant Vision Mechanisms for Resource-Constrained Humanoid Robot Applications

Rui Miguel Horta Pimentel de Figueiredo

Supervisor: Doctor Alexandre José Malheiro Bernardino
Co-Supervisor: Doctor Helder de Jesus Araújo

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering

Jury final classification: Pass with Distinction

2020

TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Space-Variant Vision Mechanisms for Resource-Constrained Humanoid Robot Applications

Rui Miguel Horta Pimentel de Figueiredo

Supervisor: Doctor Alexandre José Malheiro Bernardino
Co-Supervisor: Doctor Helder de Jesus Araújo

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering

Jury final classification: Pass with Distinction

Jury

Chairperson: Doctor Mário Alexandre Teles de Figueiredo, Instituto Superior Técnico, Universidade de Lisboa
Members of the Committee:
Doctor José Alberto Rosado dos Santos Victor, Instituto Superior Técnico, Universidade de Lisboa
Doctor Jorge dos Santos Salvador Marques, Instituto Superior Técnico, Universidade de Lisboa
Doctor Alexandre José Malheiro Bernardino, Instituto Superior Técnico, Universidade de Lisboa
Doctor Jan Paul Siebert, School of Computing Science, University of Glasgow, UK
Doctor Jorge Nuno de Almeida e Sousa Almada Lobo, Faculdade de Ciências e Tecnologia, Universidade de Coimbra

Funding Institutions:
Fundação para a Ciencia e a Tecnologia
Instituto de Sistemas e Robótica, Instituto Superior Técnico

2020

Abstract

In order to explore and understand the surrounding environment in an efficient manner, humans have developed a set of space-variant vision mechanisms that allow them to actively attend different locations in the surrounding environment and compensate for memory, neuronal transmission bandwidth and computational limitations in the brain. Similarly, humanoid robots deployed in everyday environments have limited on-board resources, and are faced with increasingly complex tasks that require interaction with objects arranged in many possible spatial configurations.

The main goal of this thesis is to assess the viability and performance of biologically inspired, space-variant human visual mechanism benefits, when combined with state-of-the-art algorithms for different visual tasks (i.e. from detection, object recognition to 3D reconstruction and pose estimation), ranging from low-level hardwired attention vision (i.e. foveal vision) to high-level visual attention mechanisms. We delve beyond the state-of-the-art in biologically plausible space-variant resource-constrained vision architectures, for active recognition and localization, volumetric reconstruction, pose estimation and multiple object tracking tasks.

Our contributions are fourfold. First, we propose an object recognition and localization framework that combines Convolutional Neural Networks (CNNs) with smooth artificial foveal vision. An iterative foveation mechanism that mimics the human saccadic eye movements, is used to improve recognition accuracy of non-centered objects within an image, over time. Second, we developed a novel a probabilistic observational model for stereo binocular systems that employes the Unscented Transform in order to propagate uncertainty in stereo matching, due to spatial quantization in the retina (represented with log-polar distributed receptive fields), to the 3D Cartesian domain. Third, we developed 3D orientation selectivity mechanisms for the incorporation of orientation-specific priors, in resource-constrained egocentric object reconstruction and allocentric pose estimation. Namely, a novel versatile structure whose topology is flexible, and may be biased according to the autonomous agent prior knowledge and short-term tasks and goals, via the allocation of limited resources to important task-dependent directions. Finally, we propose a framework inspired by divided focal attention and working memory limitations in the human brain, that poses the Multiple Object Tracking (MOT) problem as a resource-constrained decision making under uncertainty process, that keeps computations tractable by dividing and constraining visual information processing to a limited number of targets, at each time instant.

Keywords: Biologically Inspired Vision, Active Vision, Space-Variant Vision, Selective and Divided Attention, Object Detection, 3D Reconstruction

Resumo

A fim de explorar e compreender o ambiente circundante de maneira eficiente, os seres humanos desenvolveram um conjunto de mecanismos de visão variante no espaço que lhes permite atender ativamente a diferentes regiões do ambiente circundante e compensar limitações de memória, largura de banda de transmissão neuronal, e computacionais no cérebro. Da mesma forma, robôs humanoides implantados em ambientes quotidianos têm recursos limitados, e deparam-se com tarefas complexas que exigem interação com objetos dispostos numa infinidade de configurações espaciais possíveis.

O objetivo principal desta tese é avaliar a viabilidade e o desempenho de mecanismos visuais variantes no espaço com inspiração biológica no olho humano, quando combinados com algoritmos de última geração para diferentes tarefas visuais (detecção, reconhecimento, reconstrução 3D e estimativa de pose de objectos), replicando mecanismos de baixo nivel (i.e. visão foveal) com mecanismos de alto nível de atenção visual. Nesta tese, vamos além do estado da arte em arquiteturas de visão activa biologicamente inspirada, para detecção, reconstrução volumétrica, estimação de pose e seguimento de objetos. As nossas contribuições são as seguintes: Primeiro, propomos uma *framework* para detecção de objetos que combina Redes Neuronais Convolucionais com visão foveal artificial. Um mecanismo iterativo que imita os movimentos oculares sacádicos humanos, é usado para melhorar a precisão de reconhecimento de objetos não centrados no campo visual, ao longo do tempo. Em segundo lugar, desenvolvemos um novo modelo observacional probabilistico para sistemas binoculares que emprega uma transformada Unscented de forma a propagar incerteza no matching stereo no dominio da retina, devido a quantização espacial para o dominio 3D. Em terceiro lugar, desenvolvemos mecanismos de seletividade de orientação, para tarefas de procura e reconstrução de objetos em referenciais egocêntricos, assim como para estimativa de pose em referenciais alocêntricos. Nomeadamente, propomos uma estrutura para procura visual, reconstrução 3D e estimação de pose, cuja topologia é flexível e, ao contrário de outras estruturas de memória espacial existentes na literatura, pode ser enviesada de acordo com os objetivos a curto prazo e conhecimento a priori do agente autónomo. Finalmente, propomos uma arquitectura inspirada em atenção focal dividida e limitações de memória no cérebro humano, que coloca o problema de seguimento de múltiplos objectos como um processo de tomada de decisão sob incerteza, restringindo o processamento de informação visual para um número limitado de alvos, a cada instante temporal.

Palavras chave: Visão biologicamente inspirada, Visão activa, Visão variante no espaço, Mecanismos de atenção selectiva e dividída, Detecção de objectos, Reconstrução 3D, Estimação de pose, Seguimento de multiplos objectos

Acknowledgements

This thesis is dedicated to my friends at VisLab, ISR and IST, without whom it could ever have been accomplished. It was a pleasant journey, only possible with the daily friendship and help of all the ones involved. I would like to thank my supervisor and friend professor Alexandre Bernardino (Alex), who was always available for providing me with the best technical and scientifical guidance during all these years. Alex was more than a scientific advisor but also a friend, a professional and personal inspiration. I would like to thank my co-supervisor professor Helder Araújo for his scientific advice during the planning of each work, essential for easily ensuring that each scientific paper could be accepted in the top international conferences and Journals. I would like to thank professor José Santos Vitor, for the technical, scientific, and charismatic leadership. José was an inportant inspiration to pursue my master and doctoral level academic studies, as well as an early scientific career in VisLab. As the lab director, José ensured full access to state-of-the art scientific literature and equipment, and organized events, lectures, research projects with the best scientific communities. But mainly, he provided unmeasurable scientific wisdom, the character and values of friendship and sharing for free. I would like to thank all my colleagues and friends in the lab, during this journey, namely Plinio Moreno, Joao Avelino, Nuno Moutinho, Ricardo Ferreira, Ricardo Ribeiro, Atabak Dehban, Pedro Vicente, Nino Cauli, Giovanni Saponaro, Lorenzo Jamone, Nuno Monteiro, Ricardo Nunes, and all the other colleagues and friends at VisLab, ISR, IST and outside of the university, with whom I've shared the best moments and that in one way or the other were important to achieving the best scientific research outcomes. I would also like to thank professors Pedro Lima, Joao Xavier, Jorge Salvador Marques, José Santos Victor and Mário Figueiredo for sharing their valuable knowledge in their lectures. Finally, I would like to thank FCT, ISR, IST / U.Coimbra for funding my studies and the scientific work carried out during this thesis. To my family, in particular to my parents, brother, sister-in-law and niece for the emotional support provided during this stage, and finally for the most joyful moments, for the love and immeasurable affection...

to Catarina

Agradecimentos

Esta tese é dedicada aos meus amigos do VisLab, ISR e IST, sem os quais nunca poderia ter sido realizada. Foi uma viagem agradável, só possível com a amizade e ajuda diária de todos os envolvidos. Agradeço ao meu orientador e amigo professor Alexandre Bernardino (Alex), que esteve sempre disponível para me fornecer a melhor orientação técnica e científica durante todos esses anos. O Alex foi mais do que um conselheiro científico, mas também um amigo, uma inspiração profissional e pessoal. Agradeço ao meu co-orientador professor Helder Araújo pelo aconselhamento científico durante o planeamento de cada trabalho, essencial para garantir facilmente que cada artigo científico estivesse ao nível das melhores conferencias e revistas internacionais. Agradeço ao professor José Santos Vitor, pela liderança técnica e científica. O José foi e é uma inspiração importante para prosseguir os meus estudos académicos de mestrado e doutoramento, e opção por uma carreira científica, iniciada no VisLab. Como diretor do laboratório, o José garantiu acesso a equipamentos e literatura científica de última geração, além de organizar eventos, palestras, e projetos com as melhores comunidades cientificas. Mas, principalmente, transmitiu sabedoria científica incomensurável, e os valores de amizade e partilha. Gostaria de agradecer a todos os meus colegas e amigos de laboratório, durante esta viagem, nomeadamente ao Plinio Moreno, João Avelino, Nuno Moutinho, Ricardo Ferreira, Ricardo Ribeiro, Atabak Dehban, Pedro Vicente, Nino Cauli, Giovanni Saponaro, Lorenzo Jamone, Nuno Monteiro, Ricardo Nunes, e todos os outros colegas e amigos no VisLab, ISR, IST e fora da universidade com quem partilhei os melhores momentos, e que de uma forma ou de outra foram importantes para alcançar os melhores resultados científicos, com alegria e optimismo mesmo nos momentos mais adversos. Agradeço também aos professores Pedro Lima, João Xavier, Jorge Salvador Marques, José Santos Victor e Mário Figueiredo por compartilharem os seus valiosos conhecimentos nas cadeiras leccionadas. Gostaria também de agradecer à FCT, ISR, IST / U. Coimbra pelo financiamento do trabalho científico realizado durante a tese. À minha familila, em particular, aos meus pais, irmão, cunhada, e sobrinha pelo apoio emocional durante esta etapa, e por último pelos momentos mais alegres, pelo amor e carinho incomensurável...

à Catarina

Acronyms

AIP Anterior Parietal Lobe

ANN Artificial Neural Networks

BO Bayesian Optimization

CAD Computer-Aided Design

CIP Caudal Intraparietal Sulcus

CNN Convolutional Neural Network

DNN Deep Neural Networks

DCNN Deep Convolutional Neural Network

GMM Gaussian Mixture Model

GP Gaussian Proccess

GHT Generalized Hough Transform

HOG Histogram of Gradients

HRI Human Robot Interaction

HVS Human Visual System

IOR Inhibition of Return

JPDA Joint Probabilistic Data Association

LGN Lateral Geniculate Nucleus

MAB Multi-Armed Bandit

MCTS Monte Carlo Tree Search

MOT Multiple Object Tracking

NBV Next-Best-View

POMDP Partially Observable Markov Decision Process

RF Receptive Field

RPN Region Proposal Network

SES Sensory Ego-Sphere

SIFT Scale Invariant Feature Transform

UCB Upper Confidence Bound

UT Unscented Transform

WTA Winner Take All

RoI Region of Interest

FPS frames per second

MOTP Multiple Object Tracking Precision

Contents

List of Tables

List of Figures

List of Publications

P.I A. F. Almeida, R. Figueiredo, A. Bernardino, J. Santos-Victor, "Deep Networks for Human Visual Attention: A hybrid model using foveal vision", Third Iberian Robotics Conference (ROBOT), November 2017

P.II C. Melício, R. Figueiredo, A. Almeida, A. Bernardino and J. Santos-Victor, "Object detection and localization with Artificial Foveal Visual Attention", IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-EpiRob), 2018

P.III R. Figueiredo, Alexandre Bernardino, José Santos-Victor, Helder Araújo, "On the advantages of foveal mechanisms for active stereo systems in visual search tasks", Autonomous Robots, 2018

P.IV R. Figueiredo, J. Avelino, A. Dehban, A. Bernardino, P. Lima and H. Araújo, "Efficient Resource Allocation For Sparse Multiple Object Tracking", International Conference on Computer Vision Theory and Applications (VISAPP), 2017

P.V R. Figueiredo, P. Moreno and A. Bernardino, "Robust cylinder detection and pose estimation using 3D point cloud information", IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC). Coimbra, Portugal, 2017

P.VI R. Figueiredo, A. Dehban, A. Bernardino, J. Santos-Victor, H. Araújo, "Shape-Based Attention for Identification and Localization of Cylindrical Objects", IEEE International Conference on Developmental and Learning and on Epigenetic Robotics (ICDL-EpiRob). Lisbon, Portugal, 2017.

P.VII R. Figueiredo, A. Dehban, P. Moreno, A. Bernardino, J. Santos-Victor, H. Araújo, "A Robust and Efficient Framework for Fast Cylinder Detection", Robotics and Autonomous Systems (RAS), July 2019

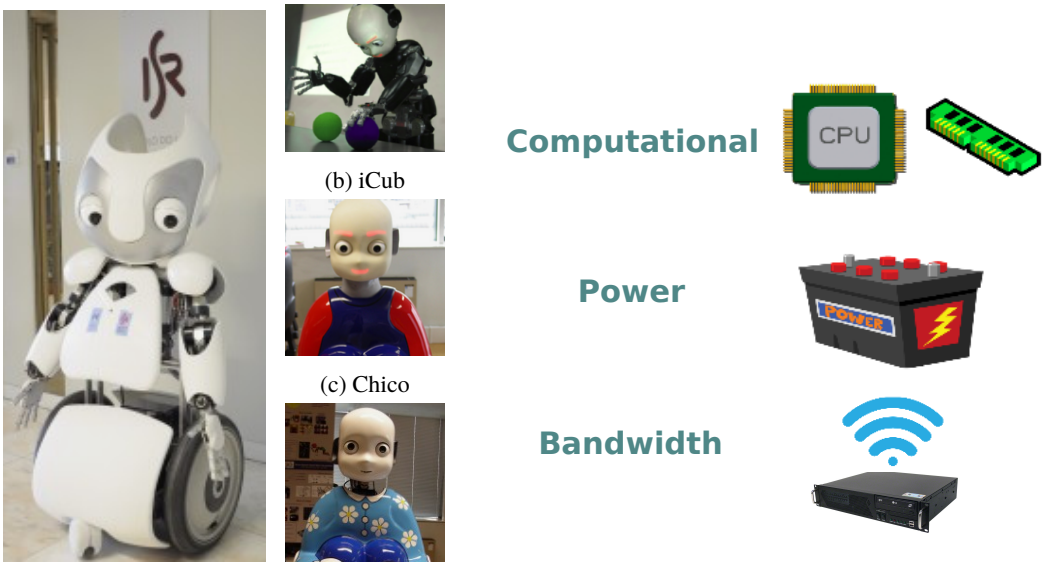
Chapter 1

Introduction

The present doctoral dissertation aims at improving the state-of-the-art in biologically inspired computer vision solutions for robotics visual tasks,, including active recognition, volumetric reconstruction, pose estimation and multiple object tracking. More specifically, this thesis studies and models space-variant visual attention and constrained resource allocation mechanisms existent in humans, showing, in a set of visual tasks relevant to robotics, the advantages and disadvantages of these mechanisms in comparison with the classic and modern computer vision approaches. Unlike other studies that focused on developing hardware based retinas, we focused on software based ones.

1.1 Motivation

Space-variant vision and attention mechanisms are the fundamental processes in biological systems, responsible for prioritizing the elements of the visual scene to be attended, i.e., to control perceptual resources [6, 158] and cope with the brain computational limitations. Humans rely on space-variant sensing (foveal vision), and on stimulus-driven (bottom-up) and goal-driven (top-down) information processing mechanisms to define where in the visual input the attentional foci should be oriented to [109]. This way, information processing is constrained and directed towards salient or task-relevant stimuli.
Likewise, an important issue in many computer vision applications requiring real-time performance, resides in the involved computational effort [26], especially in robotics where energy efficient, fast and accurate perception is a fundamental requirement, e.g., in visual localization and servoing during grasping, manipulation and hand-over of tools to human or machine collaborators. In humanoid robotics, in particular, real-time operation is conditioned by physical limitations on on-board computational and power resources, as well as data transmission bandwidth if one opts to outsource information processing to outside servers (see Figure 1.1).



(a) Vizzy (b) iCub (c) Chico (d) Chica (e) Resource Constraints

Figure 1.1: Humanoid robotic platforms available at VisLab (http://vislab.isr.ist.utl.pt/) (Left) and their main resource limitations (Right)

Therefore much effort has been made towards understanding the underlying principles of biological attention mechanisms and applying those mechanisms in robotics, in an attempt to build more efficient solutions, capable of performing in real-time, under resource-constrained settings [17].

In the remainder of this chapter we overview the neurophysiology of the Human Visual System (HVS), and review the state-of-the-art in biologically plausible space-variant vision models, focusing on artificial foveal vision and visual attention mechanisms.

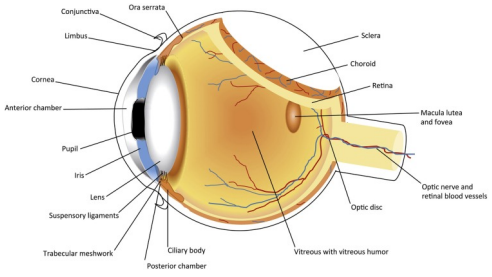
1.2 Neural and Artificial Mechanisms of Visual Information Processing

The process of seeing starts with light entering the eye through the cornea. The eye has the ability to adapt to different levels of brightness (adaptation) and to shape its lens and pupil size in order to focus objects at different distances (accommodation). The light passing through the pupil, is focused by the lens, onto the retina, a sensory membrane responsible for receiving and converting the visual stimuli into electric signals to be transmitted to the visual cortex in the brain through the optic nerve [143].

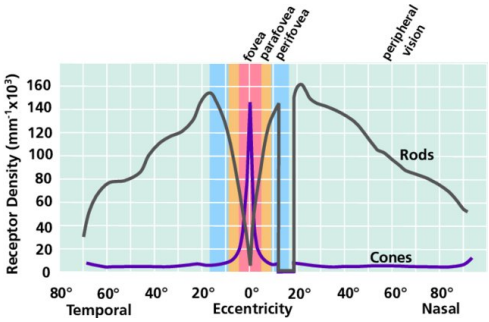
The retina is mainly composed of two types of photo-receptors: rods which are mostly concentrated, at the periphery and are sensitive to brightness and colorless low-light vision (scotopic vision) and the cones that are concentrated mostly in the center of the eye, in a place called fovea, and are responsible for high acuity color vision (see Figure 1.2). Finally, the visual signals entering through the optic nerve reach the back of the brain, where the visual cortex is located and the stimuli interpreted.

1.2.1 Space-variant Foveal Vision

Unlike uniform vision provided by conventional imaging sensors, human vision is space-variant, due to the uneven organization of the photo-receptors in the retina. Visual acuity, provided by the cones, is highest at the fovea, located in the center of the retina, and declines monotonically towards the periphery, with increasing eccentricity (see Figure 1.2). This space-variant resolution perception phenomenon - named foveation - is a hardwired mechanism and a natural way of reducing the amount of information streamed to the brain, in order to cope with power, neuronal transmission bandwidth limitations, and the brain machinery processing capacity. In fact, if foveal resolution visual stimuli across the whole field of view was to be processed, the human brain weigh would be significantly increased (to approximately 60 kg [14]). However this compression phenomenon introduces a space-variant uncertainty in visual processes. In order to efficiently explore and understand the surrounding environment [162], humans have developed a set of attention and oculo-motor mechanisms, namely saccades, that allow them to actively and sequentially direct their eyes towards different regions of interest in the surrounding environment, and thus, to cleverly compensate for the aforementioned limitations.



(a) Eye morphology (figure taken from [143])



(b) Photo-receptors density in the retina (figure taken from [215])

Figure 1.2: Human Eye Physiology

Similar to humans, robots deployed in everyday environments, are faced with increasingly complex scenarios where objects are arranged in many possible different spatial configurations. The problem of deciding which regions in the visual field are to be attended during visual search tasks is computationally demanding or even intractable if approximate solutions are not considered [206]. Therefore, like biological systems, humanoid robots must be endowed with mechanisms to allow them to locate objects of interest and to sequentially build detailed representations of the scene, while avoiding the potential overload of processing irrelevant sensory stimuli. Under the assumption that biological systems perform quasi-optimally in their environment due to multiple generations of

genetic improvement, researchers have been developing robotics systems [135] provided with biologically inspired space-variant image processing [184, 202], gaze control models [175, 21] and attention systems [17, 26, 210]. These implementations not only mimic the mechanisms observed in humans but, in general, also lead to more efficient and effective behaviors under resource-constrained settings (bandwidth, computational and energetic). In the context of robotics, and from a practical standpoint, unconventional space-variant sensing representations, in particular human-like foveal vision, offer multiple advantages when compared to conventional uniform counterparts, including reduced resolution with wide field-of-view, being suitable for real-time performance in active vision systems that are able to manipulate the view-point.

1.2.2 Computational Foveal Vision Mechanisms

All levels of the visual system are highly regular and symmetric, from the photoreceptors distribution in the retina, to higher-level cell organization in the striate cortex. Different digital sensing architectures exist in the literature that attempt to mimic biological vision structures, namely adaptive and reconfigurable hardware-based ones [67, 12], in this thesis we focus on algorithmic-based human like vision ones.

Biologically plausible foveated digital image processing techniques attempt to mimic the space-variant phenomena in the visual pathways, and have numerous applications, including video streaming in low-bandwidth networks (e.g. teleoperation and remote surveillance) and scene understanding tasks (e.g. object detection [5], tracking [21, 74], and robot navigation [181]). The algorithms proposed in the literature, try to mimic foveal vision and can be classified as geometric [202], multi-resolution [1], or filtering-based [68, 214].

Geometric-based Approaches

Studies from neurophysiology have shown that the receptive field spacing and size scale exponentially with eccentricity in the retina, and that light stimuli produces activation displacements in the cortex that are inversely proportional to the distance to the fovea.

Geometric-based approaches attempt to model the retinotopic mapping transformation, that occurs between RFs in the retina and the Lateral Geniculate Nucleus (LGN) [95], where neighboring retinal locations are mapped to neighboring cortical locations. This RFs mapping distribution can be mathematically approximated using the log-polar transformation [183], which is given by the following mathematical expression:

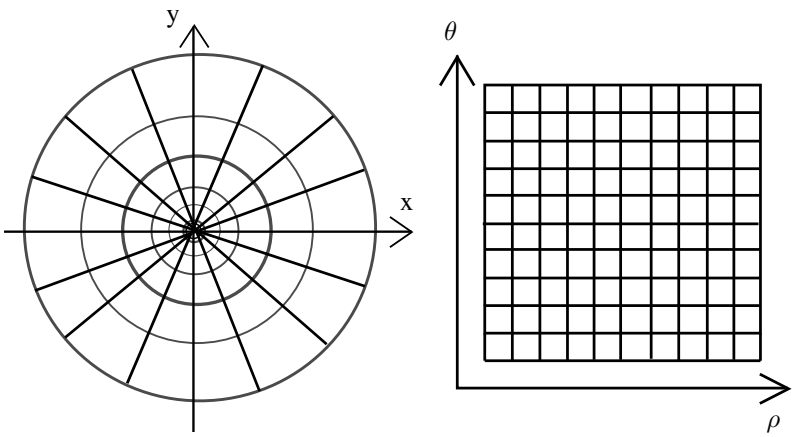
$$(\rho, \theta) = \left(log\left(\sqrt{((x - x_c)^2 - (y - y_c)^2)} \right), atan\left(\frac{(y - y_c)}{(x - x_c)} \right) \right) \tag{1.1}$$

and has attracted much interest within the robotics community (see Figure 1.3a (and Figure 1.3b). First, because it allows trading-off field-of-view, resolution and data compression. Second, they provide some degree of invariance to rotations and scaling transformations, as these become linear shifts in the cortical plane.

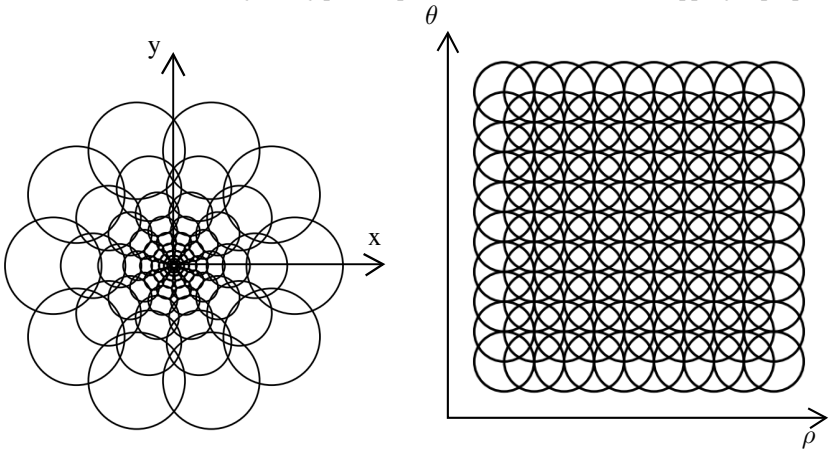
Many log-polar models have been proposed in the literature [24] and may be categorized as conformal non-overlapping or overlapping, depending on the RF support radius (see Figure 1.3). Although being computationally more intensive than their non-overlapping RFs counterparts, overlapping models are better at approximating the space-variant averaging phenonema in the retina, and produce smoother retinal mappings. Still, the literature falls short on works that attempt to model uncertainty in 3D reconstruction due to space-variant quantization phenomena in the retina, and to leverage these uncertainty measures for Next-Best-View (NBV) planning during exploration and visual search tasks. This is one of the contributions of the thesis, described in chapter 3.

While previous approaches attempt to capture the retina receptive field tessellation structure through analytic geometric modeling, other approaches capture its underlying structure through exploration and learning strategies. One example is the self-organized retina of [13] that unlike previous approaches can deal with sampling discontinuities between the fovea and the peripheral region of the visual field. During the structure creation process,

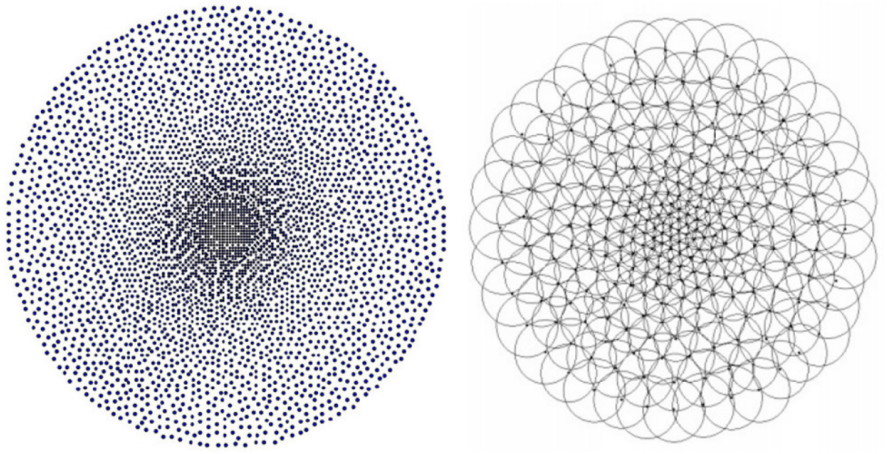
they use self-similar neural network units, whose weights undergo random transformations to produce randomized tessalations (see Figure 1.3c)



(a) Retinal (left) and cortical (right) log polar representations with non-overlapping superpixel RFs.



(b) Retinal (left) and cortical (right) log-polar representations with overlapping circular RFs. Left: the x and y correspond to Cartesian coordinates in the retinal plane. While ρ and θ correspond to coordinates in the cortical domain.



(c) Self-organized Gaussian receptive field tessellation produced with self-similar neural network units. Left: node tessellation. Right: Gaussian receptive fields on top of a retina tessellation.

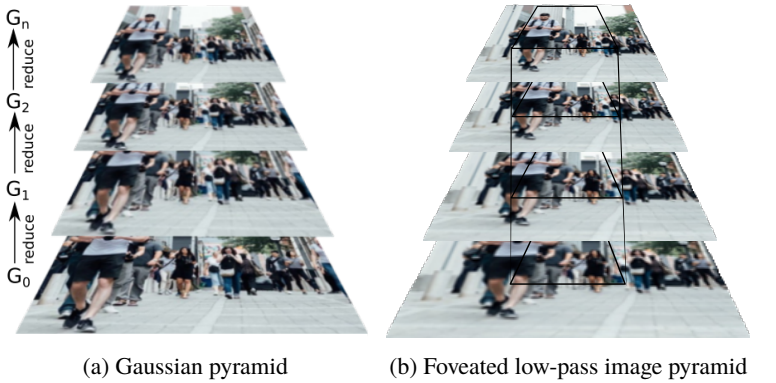
Figure 1.3: Log-polar transform.

Multi-resolution Pyramids

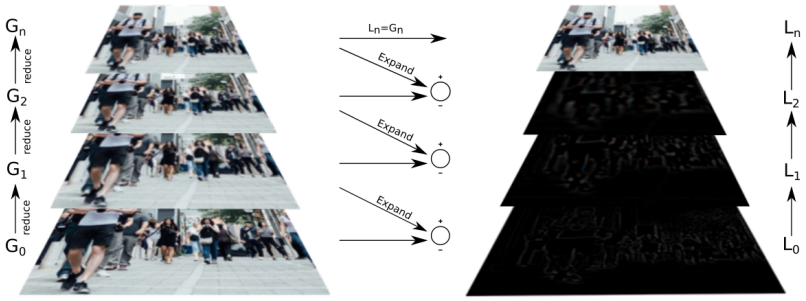
Image pyramids [1] have been proposed for multi-resolution image processing and are particularly suited for multi-scale image analysis, data compression, and as an intermediate step of key point extraction algorithms (e.g. Scale Invariant Feature Transform (SIFT)). The basic principle resides on low-pass smoothing and downsampling.

Gaussian pyramids are the most common in the literature and utilize Gaussian kernels for the smoothing operation. The first level in the pyramid (level 0) contains the original image g_0 that is first low-pass filtered via convolution with $2D$ isotropic and separable Gaussian filter kernels, and then downsampled by a factor of two, yielding the image g_1 at level 1. Successive images g_{k+1} are obtained from the previous levels g_k, by iteratively repeating the low-pass filtering and down-sampling procedures (see Figure 1.4a). Gaussian pyramids are useful for many applications, in particular for recognizing patterns of unknown scale (e.g. scale invariant template matching), and for fast foveated coarse-to-fine pattern localization (see Figure 1.4b).

The Laplacian pyramid (see Figure 1.4c) was first introduced in [32], for image compression, and is constructed by computing differences of Gaussians. During the construction of the pyramid, each level of the Gaussian pyramid g_k is subtracted from an expanded version of g_{k+1}, to ensure similar resolution and obtain a band-pass image L_k. Data compression is achieved by storing the largely decorrelated L_k and the low-pass filtered image g_{k+1}, instead of g_k.



(a) Gaussian pyramid (b) Foveated low-pass image pyramid



(c) Laplacian pyramid

Figure 1.4: Multi-resolution Pyramid Representations

Filtering-based Methods

In the work of [68] the authors proposed a foveation mechanism for digital image streaming in low-bandwidth communication channels, that allows the user to point the high spatial resolution focus to regions of interest, with pointing devices (e.g. eye tracker or mouse), being also suitable for studies involving eye movements. The method starts by building a Laplacian pyramid, then, each level is multiplied by an exponential kernel, centered at the foveation point, upsampled and summed with the previous levels, to obtain an image that matches the psychophysical space-variant contrast sensitivity of the HVS (see Figure 1.5). Matching the falloff resolution of the HVS, makes optimal use of compression resources, by discarding only the details that cannot be resolved by the human eye, via manipulation of the exponential kernel standard deviation.

Inspired by this model we developed a real-time implementation that was used to study the impact of artificial foveal vision mechanisms in gaze sequence modelling. The contribution is overviewed in chapter 2.

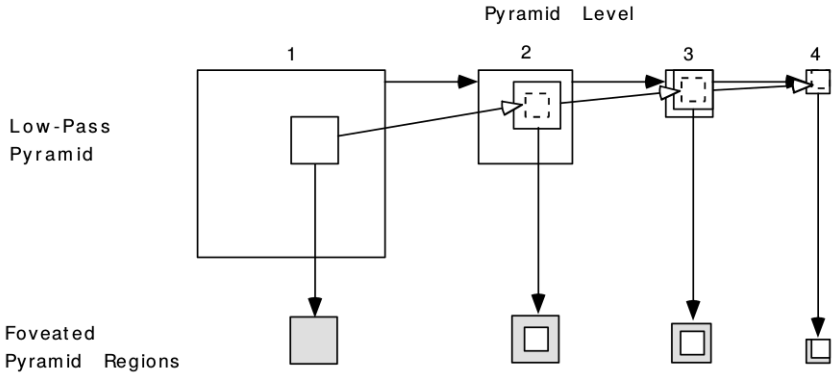


Figure 1.5: Filtering-based foveation (figure adapted from [68]).

1.2.3 Visual Attention and Spatial Selectivity as Resource Constrained Perception

Visual attention is the process through which organisms select a sub-part of the visual stimuli to be processed in detail, while suppressing the rest, to obtain an efficient perception and cope with limited brain computational resources.

The first studies on visual attention date back to the mid 19th century, pioneered by Hermann Von Helmholtz [211] and motivated by the willingness to understand how humans attend stimuli at the periphery of the visual field. By designing a device called tachistoscope Helmholtz demonstrated independence between the ocular attention focus (i.e. gaze location) and the peripheral attentional foci.

The first model for visual attention was proposed by Broadbent [28, 167], in his filter theory, which introduced the structural bottleneck concept (a limitation on the amount of information that the brain can process), that suggests that selective filters are necessary to decide which stimuli to process and which to ignore. Nowadays, the literature on visual attention is vast, and covers a wide range of scientific fields, including cognitive neuroscience [35] and computer science [25], playing an important role in computer vision and robotics applications [17]. Attention modeling is not just a multidisciplinary but also a challenging topic under active research due to its importance in controlling the regions (where) and the features or objects (what) the observer should attend to, over time (when). Attention mechanisms can be either selective or divided.

Seminal studies from Hubel and Wiesel [96, 97] suggest that the RFs in the mammalian visual cortex increase in size along the visual stream, covering wider areas of the visual field. In parallel, information is selectively processed and the abstraction level of the representations selected along the visual pathways, increase in complexity and in a hierarchical tree manner. Selective attention mechanisms deploy resources to single features or locations, in a serial manner, while divided mechanisms prioritize resources to multiple features or locations, in a parallel

manner.

Selective Attention Mechanisms

Selective visual attention mechanisms are the processes through which biological organisms select only part of the visual signal to be processed while suppressing and ignoring the rest to obtain an efficient perception, and cope with limited neural resources in the brain, allocated to vision. It covers all factors that influence information selection mechanisms, whether they are driven by visual stimuli (bottom-up) or by task-related expectations (top-down) [23]. In particular, spatial attention has been often compared to a spotlight that selectively discards information outside a subarea of the field-of-view. The more sophisticated zoom lens model of [50] suggests that the size of the attentional spotlight is dynamic and object magnification inversely proportional to the lens power (i.e. the spotlight size).

Other selective attention theories attempt to explain feature integration [204], based on [203] the idea of determining which visual features are detected preattentively and how the visual system makes the preattentive processing [204]. To identify the preattentive features, [204] made experiments to detect targets and measuring performance response time and accuracy. In the response time model, viewers were asked to complete the task as quickly as possible and the number of distractors on the display varied. To understand how preattentive processing is done, Treisman proposed a model (see Figure 1.6). where each feature map registers the activity of a specific visual feature channel like contrast or size. When an image is shown, features are encoded in parallel into their respective maps. These maps only provide us the activity log of each feature. If the target has a unique feature, we just have to check if there is activity on the respective feature map. However, for conjunction target, one feature map is not enough. Thereby, a serial search must be done in order to find the target that has the correct combination of features. In this case, a focus of attention is used to increase the time and effort spent.

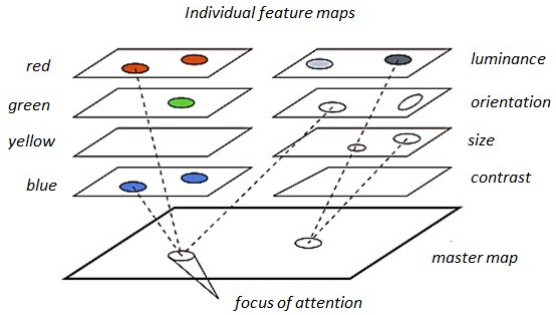


Figure 1.6: Treisman's feature integration model of early vision — detection of activity in individual feature maps can be done in parallel, but to search for a combination of features, attention must be focused. Figure adapted from [80].

Ungerleider and Mishkin [140] proposed that the visual pathways can be functionally distinguished between *ventral* and *dorsal*, both originating in the primary visual area (V1) (see Figure 1.7). The ventral stream mediates feature extraction and object recognition (what) whereas the dorsal stream is specialized in motion and location selectivity (where).

Recognition Pathway Visual stimuli entering the ventral pathway is foveal and neurons within the ventral stream respond selectively to visual features that are important for recognition tasks. Input is grouped in increasingly complex and meaningful visual elements along the pathway. Stimuli selectivity ranges from low-level orientation and color contrast selectivity in V1 and V2, to aggregated contour features and complex shapes in V4 ending in higher-level object representations in the inferior temporal (IT) cortex, which comprise category-specific cells. Visual representations are encoded in allocentric or object-centric reference frames. Neurons involved in low-level detection of disparity, were mainly found in the visual cortex, in areas V1, V2 and V3 [208], whereas neurons

involved in high-level disparity processing facilitate computation of view-point invariant object-specific attributes, to ease recognition functions.

Localization Pathway Neural circuits in the dorsal pathways are tuned for spatial location and motion detection, playing an important role in visuomotor coordination (e.g. in visually guided reaching and grasping). The dorsal stream processes both foveal and peripheral stimulus, and builds a detailed spatial map of object locations and orientations in the field of view. High-level disparity processing, or the reconstruction of 3D surface orientation through the computation of disparity gradients, were found mainly in the Caudal Intraparietal Sulcus (CIP), in the dorsal stream.

In [176], the authors studied how 3D shape orientation is visually encoded in the brain. In particular, they developed analytical methods to study neural encoding of 3D surface orientation features in the CIP, in the dorsal stream. By varying the orientation of a planar chess pattern positioned frontoparallel with respect to human subjects, the authors concluded that neurons in the CIP jointly encode pan and tilt orientation of 3D surfaces, and that the distribution of preferences over orientations is statistically close to uniform. Nevertheless, it is still unclear if other areas in the brain exhibit unbiased activation selectivity. It is known, however, that areas such as V4 are tuned for specific 3D orientations [84], and that 3D features for grasping and manipulation are context-dependent in the CIP area.

At last, although different neuro computational models have been proposed in the computer vision literature for orientation selectivity in 2D (orientation, motion), it is scarce on works that attempt to model space-variant biases for stimuli selectivity in 3D for enhanced pose estimation, which is one of the contributions of the thesis described in chapter 4.

Divided Attention Mechanisms

In divided attention, the focus of attention is split between multiple stimuli at a time. Computational resources are limited by a cognitive budget, and therefore attention mechanisms demand the separation of resources among different tasks. Early psychology studies from George A. Miller summarized evidence that the number of objects an average subject can hold in short-term memory is around ± 7, and is occasionally referred to as Miller's law [139], representing a constraint on the humans's capacity for holding objects in short-term and working memory.

Multi-focal parallel deployment of attention has been mostly associated to Multiple Object Tracking (MOT) that deals with the problem of maintaining the location and identity of multiple dynamic targets. Pylyshin and Storm [164] were the first to study humans ability to track multiple targets. They designed an experiment in which 10 identical and independent moving targets are exhibited in a multi-element display, and subjects asked to track a smaller subset, pre-defined at the begining of the trial. During the trial, all the items move randomly and independently, and in the end subjects are asked which targets were selected at the begining of the trial. They concluded that, unlike other aspects of attention that require serial scanning of the visual scene (e.g. visual search), MOT involves parallel constrained processing that sustains spatial resource allocation to locations of around four moving targets simultaneously. This finding is consistent with the magical number four in short-term memory theory which in constrast to [139], claims that an average individual can keep 4 items in the visual short-term working memory [39]. Yantis and Jones [222] refined this theory by showing with supported experimental evidence, that the limited number of attended targets, is dynamic and temporally modulated by priorities that depend on the nature of the task.

Other theories suggest that MOT is not constrained to a discrete limited number of indices, but is rather an analog phenomenon [36], in which the limited size attentional spotlight is divided between multiple spatial regions, and performance affected by bottom-up influences such as targets' proximity (crowding) and perceptual limitations including visual acuity and uncertainty on target's locations [61].

Neurological findings using fMRI data state, however, that the location of moving objects is represented in

the Anterior Parietal Lobe (AIP) and the human motion brain area (V5) [92]. In this thesis we developed a MOT algorithm that emulates the resource limitations of the human brain, in the number of attended objects and area of the field-of-view at each time instant; and the proposed approach is thorougly described in chapter 5.

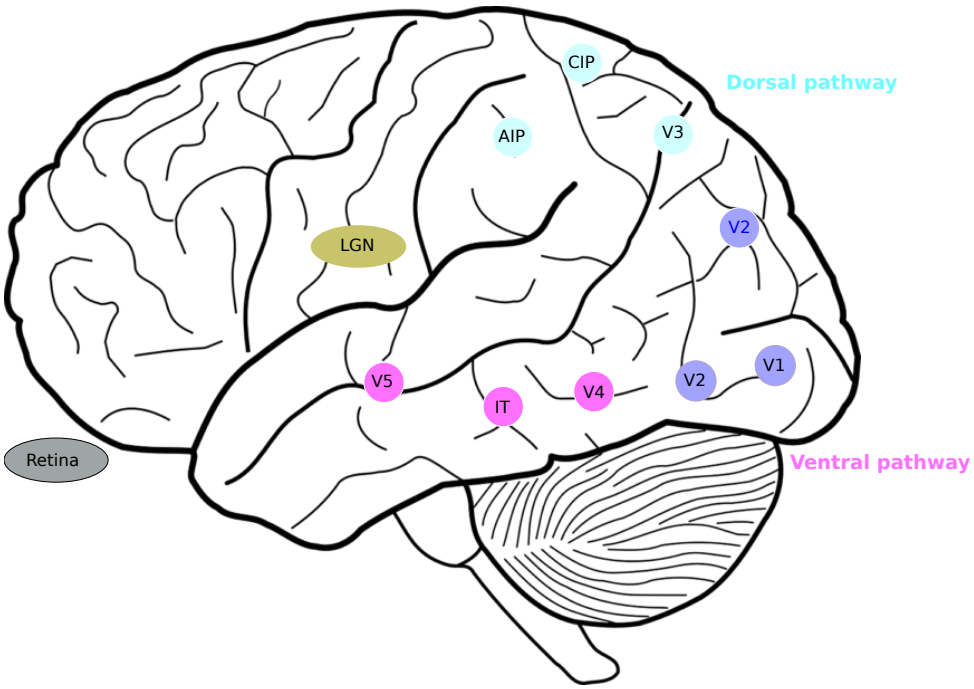


Figure 1.7: Human Visual System Pathways

1.2.4 Computational Models of Visual Attention

William James [105] defined two modes of attention orienting that facilitate the processing and selection of information: stimuli-driven (exogenous) and task-driven (endogenous). The observer attention can be stimuli-driven, triggered by scene characteristics like color or orientation (bottom-up factors) or by specific visual characteristics that depend on the current task or goal (top-down factors). On the one hand, bottom-up processing refers to the involuntary mechanisms responsible for directing resources to salient regions based on differences from a region and its surround (e.g. contrast). In this case, the stimuli directly triggers our attention and, thus, it is a data-driven process. The exogenous system is responsible for orienting our attention, in an involuntary and reflexive manner, to salient locations, features or to where sudden changes occur. For instance, when a light source flashes, ones reaction will be to reflexively direct the gaze to the source [191]. On the other hand, top-down processing corresponds to allocating attention voluntarily to features, objects or spatial regions based on prior knowledge and the agent current goals [163]. Thus, prior knowledge and the task at hand are used to influence attention in a goal-driven manner. The endogenous mechanisms are voluntary and responsible for directing the attentional resources to predetermined locations, features or objects. Orienting of attention results from taking into account task-specific internal goals, for example, when searching for specific objects or counting how many people will pass through a door. By guiding our attention to task-relevant places we make the counting process more efficient. Computational models of visual attention attempt to mimic the behavioral aspects of the HVS. The proposed models in the literature may belong to three different branches namely bottom-up, top-down, or hybrid models combining the previously.

Bottom-up Bottom-up mechanisms are agnostic to the task at hand and have the purpose of extracting relevant low-level features and finding the most salient regions where attention should be deployed.

The pioneering works of Itti [111, 102] combine multi-scale low-level features into a single saliency map. At first, spatial feature maps are built by extracting prominent local features from different feature modalities (color, intensity, orientation), using center-surround operations at different scales. Then, each map is normalized and linearly combined in a single saliency map. Finally, the Winner Take All (WTA) principle is applied to select the most salient locations to be sequentially analyzed, in order of decreasing conspicuity, using an Inhibition of Return (IOR) mechanism [199].

Osberger's approach [154] starts by performing image segmentation and then assigning perceptual importance based on low-level image features - contrast, size, shape, color and motion - and high-level features - location, people and context. Osberger chose only 5 features to use in his algorithm and, per region, assigns an importance score to each. Lastly, a combination of these features results in a map which represents important regions in an image. Kadir *et al.* [108] identify salient regions based on entropy measures of image intensity while Gao [66] defined a salient region considering how different this is from the surrounding background (center-surround mechanism [187]).

Top-down The top-down models take into account the observer's prior knowledge, expectations and current goals. The literature on visual attention suggests several sources of top-down influences [26] when the problem is to decide which stimuli is important: attention can be drawn to specific object visual features in search models to easily reach the goal or use the context or gist to constrain search locations. Whenever there is a search task, top-down processes tend to dominate guidance and target-specific features are an essential source to draw attention more effectively. Moreover, our attention is oriented to task-relevant features. This way, attentional resources are not wasted and time and computational effort are saved for processing more pertinent/relevant parts of the visual field. Under these conditions, one knows what is looking for (goal) and we know from a priori knowledge to distinguish the features that we should be searching for. Thereby as defended by guided search theory [218] [219], we are able to modulate the gains assigned to different features. If, for example, the task is to find a green object, the gain assigned to green color will be higher.

Hybrid Most visual attention approaches, model bottom-up and top-down processes independently. However, there must be a trade-off between purely bottom-up models that typically miss to detect inconspicuous objects of interest and top-down systems that confine object search according to prior expectations related to the task.

In recent years, a combination of bottom-up and top-down models, that we designate as hybrid models, have been presented. For instance, Frintrop's model [63] is compound by two saliency maps: one corresponding to top-down influences and another related with bottom-up influences. The aggregated saliency map is computed as a linear combination of those maps using a fixed weight which revealed to be a non-flexible approach. Rasolzadeh *et al.* [170] presented a more flexible model where the combination of top-down and bottom-up saliency maps is done dynamically, using entropy measures that provide information of how the linear combination of weights should change over time. Conspicuity maps were created following Itti's approach in [103] besides the extra parameters used to weight the saliency map. They used a neural network to learn the bias of the top-down saliency map based on information provided by contextual scene and the current task. These hybrid models suggest that the HVSs can guide attention by applying top-down weights on bottom-up saliency maps allowing quicker target detections in backgrounds full of distractors [170].

The authors in [224] proposed a probabilistic Bayesian framework for saliency learning using natural statistics (SUN). The most salient features are the ones with the highest point-wise self-information from features prior learned from a set of natural images, i.e., features that mostly differ from the learned average and are statistically unexpected (bottom-up modulation), or have the highest mutual information when searching for a specified target object (top-down modulation).

1.3 Objectives

The main goal of this dissertation is to study and to develop resource efficient computer vision algorithms for autonomous robotics agents that are constrained by limited on-board resources and endowed with the ability of manipulating the parameters of the visual sensor or algorithms to gracefully trade-off computational resources with performance. By borrowing ideas from cognitive neuroscience and from neurophysiology literature, the present dissertation intends to improve the state-of-the-art in resource-constrained active vision algorithms for robotics, from the sensory level to higher-level cognitive functionalities. We focus on space-variant vision phenomena, not only on the classical interpretation on the image plane (foveal vision and divided attention) but also at the level of non-uniform representations of 3D space and orientation space. The main research goals can be enumerated as follows:

Space-variant low-level vision To assess the performance advantages of low-level foveal vision sensing architectures when compared with conventional uniform Cartesian ones on recognition and visual search tasks. More specifically, to study the characteristics of low-level foveal vision mechanisms in humans, with the goal of developing more efficient artificial foveal vision models inspired by non-uniform space-variant vision phenomena in the early visual pathways of the HVS, while establishing formal relationships between computational savings and performance accuracy on recognition, detection and 3D reconstruction tasks.

Space-variant orientation selectivity

- **Space-variant selectivity models** To investigate if high-level orientation selectivity mechanisms in the visual cortex can be used as priors for enhanced visual search and pose estimation in 3D. More specifically, to develop algorithms inspired by 3D orientation selectivity mechanisms in the visual cortex, with the goal of enhancing the state-of-the-art on NBV planning and object pose estimation approaches, under resource-constrained settings. The developed mechanisms should allow incorporating environment and task-dependent

priors, with the goal of improving the accuracy in visual search and pose estimation tasks.

- **Space-variant divided attention:** To develop resource-constrained MOT algorithms, inspired by divided attention mechanisms and working memory limitations in the brain, with the goal of reducing the computational complexity of existing solutions in the literature.

1.4 Main Contributions

The main contributions of the dissertation illustrated in Figure 1.8, towards achieving the objectives enumerated in 1.3 can be summarized as follows:

1. an implementation of a biologically plausible framework for object detection in 2D, that combines low-level foveal vision with state-of-the-art Deep Convolutional Neural Networks (DCNNs).

2. a novel 3D orientation selectivity mechanism for the incorporation of orientation-specific priors, implemented using GMM for egocentric object search and allocentric object pose estimation.

3. a probabilistic observational model for stereo systems that relies on the Unscented Transform in order to propagate uncertainty in stereo matching, due to spatial quantization in the retina (represented with log-polar distributed receptive fields), to the 3D Cartesian domain

4. a framework that poses the MOT problem as a resource-constrained Partially Observable Markov Decision Process (POMDP), and keeps computations tractable by limiting the number of targets attended, at each time instant, belonging to specific image subregions.

Throughout the rest of the thesis we overview in detail each contribution.

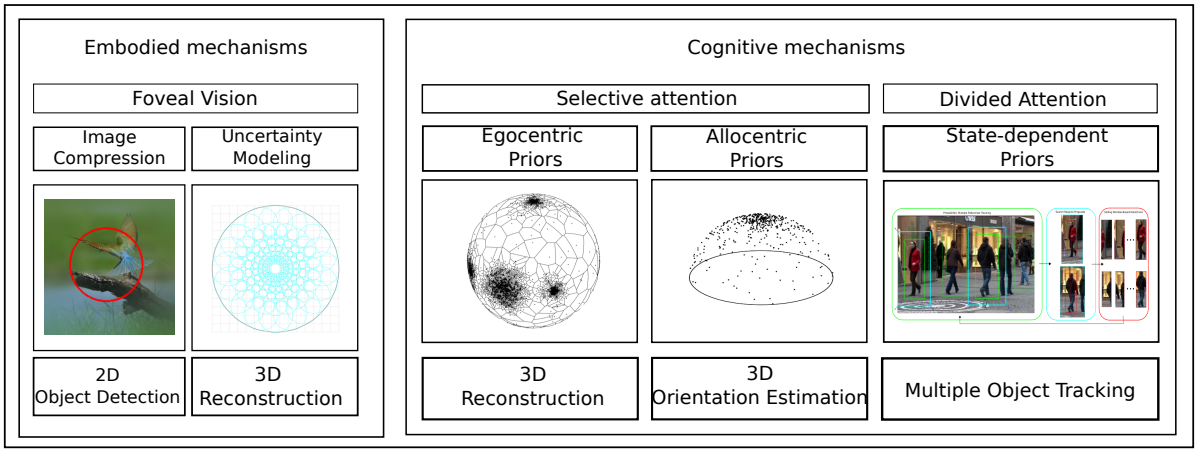


Figure 1.8: Proposed Space-Variant Vision Mechanisms

1.5 Outline of the thesis

The remainder of this document is organized as follows. Chapter 2 presents a novel hybrid model object classification and localization framework that combines feedback and feed-forward mechanisms using Convolutional Neural Networks (CNNs) and smooth artificial foveal vision, with the goal of studying the trade-off between performance and accuracy obtained with different non-uniform foveal strategies. In chapter 3 we propose a novel visual search framework for robotic systems provided with binocular foveal vision with the goal of studying the advantages of

foveal mechanisms, sampled according to the log-polar transformation, when compared with conventional uniform ones.

In chapter 4, we study the benefits of incorporating orientation selectivity priors in 6D pose estimation using 3D point cloud information. These benefits are addressed in the problem of detecting cylindrical shapes, that are commonly found in household and industrial environments. Chapter 5 introduces a novel probabilistic framework which poses the MOT problem as an on-line, resource-constrained decision making problem. The tracking constraints are inspired by divided attention mechanisms in the brain. In particular, limitations in the number of targets and size of the attentional focci. Finally, in chapter 6, we summarize our main contributions and suggest ideas for future work.

Chapter 2

Object Detection with Smooth Foveal Vision

This chapter goal is to assess how foveal vision mechanisms affect the performance of state-of-the-art deep learning approaches for object classification tasks, and investigate novel ways of compensating low visual spatial acuity with artificial saccadic mechanisms. A set of experiments demonstrate that human-like foveal vision is more efficient and compare its effectiveness on visual recognition tasks. More specifically, in this chapter we propose a biologically inspired object classification and localization framework that combines DCNN with foveal vision. We study ways of localizing objects in foveated images, where objects may lie in random locations in the visual field.

2.1 Introduction

In visual recognition tasks humans perform frequent saccadic movements in search of relevant items. This search is not random, but guided by space elements that suggest the presence of certain features or objects. In this direction we aim to develop methods that perform guided visual search tasks and not just random ones. The visual information obtained during search tasks should suggest the presence of possible objects in the periphery of the visual field that should guide the following saccades.

The work described in this chapter is inspired by the work of Cao *et al.* [34] that proposed to capture visual attention through feedback DCNNs. The method called *Look and Think Twice* is utilized to locate an object, in a top-down manner. He utilizes feedback CNNs pre-trained to classify objects from the ImageNet dataset comprising more than 1000 classes, and performs two passages through the network. In the first feed-forward pass, the predicted class labels are obtained, providing a notion of the most probable object classes that are presented in the input image. Then, based on the top-ranked labels given by the network, the method extracts the saliency map of the image with respect to each one of the top-5 class labels. In the second feed-forward pass, the original image is cropped around the salient region, and re-classified, providing a new set of predicted class labels. The classification results at the second pass are typically more accurate than at the first sight.

Similarly in spirit, we propose a biologically inspired hybrid attention model, that is capable of efficiently recognizing and localizing objects in digital images, using human-like foveal vision. Our method also applies two passes on the image, the first coarsely classifies which object is present in the field-of-view and the second localizes where these objects may lie. However, we use a foveal image representation and, instead of cropping the salient region, we simulate an actual saccade to redirect the fovea to the center. Our method also applies two passes on the image, the first coarsely classifies which object is present in the field-of-view and the second localizes the region where these objects may lie. However, we use a foveal image representation and, instead of cropping the salient region, we simulate an actual saccade to redirect the fovea to the center.

The main contributions of this work are the following: first, we establish a formal relationship between performance and the different levels of information preserved by foveal sensing configurations. Then, we evaluate the performance of our methodology for various well-known CNN architectures that are part of the state-of-the-art in detection and localization of objects when combined with multi-resolution, human-inspired, foveal vision. The remainder of this chapter is organized as follows: In section 2.2 we review the main concepts behind object detection with an emphasis on deep learning approaches, in section 2.3 we describe in detail the proposed methodologies, including a saliency-based selective attention mechanism for class-specific object localization. In section 2.4, we quantitatively evaluate our contributions, and finally, in section 2.5, we wrap up with conclusions and draw ideas for future work.

2.2 Theoretical Background

Object classification consists of assigning a single label to a given image. Localisation includes not only classifying the subject of an image but also identifying its position, usually by means of a rectangular bounding box. Object detection assumes the possibility that more than a single instance can exist in a single image, namely of different classes. Thus the desired output consists of every instance's class label and respective bounding box.

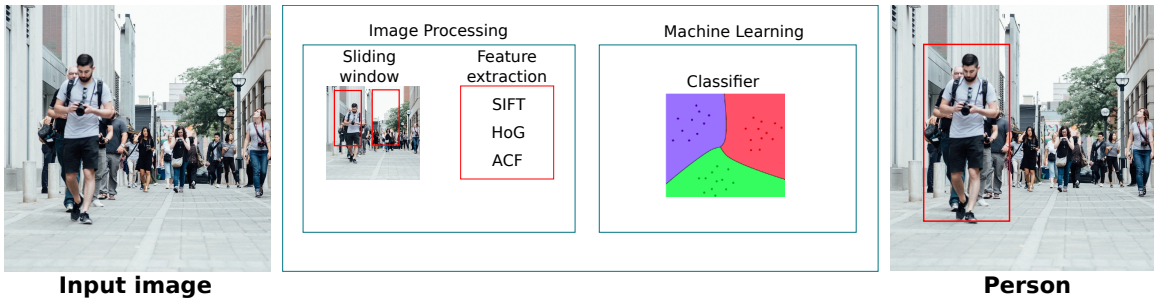
Classical methods for visual recognition tasks in the computer vision literature, extract key point features from the image, using hand-crafted filters, namely Histogram of Gradients (HOG) [42] or SIFT [128]. During a training phase, features are extracted from a set of different viewpoints, and stored in a database. In the online recognition phase, extracted features are matched against the database, based on their Euclidean distance. The implementation is typically a hash table and the Generalized Hough Transform (GHT) employed for fast and robust model matching. One successful example in the literature is the Aggregated Channel Features (ACF) of [46] for pedestrian detection, which employs a sliding window detection by classification approach, in which each window is binary classified as "person" or "not a person". Classification is performed using boosted decision trees, trained with labeled samples of full body pedestrians, using the Adaboost algorithm [62]. The classification method relies on handcrafted features that combine several image channels: LUV, Gradient Magnitude and HOGs channels aggregated in a blockwise manner. For multi-scale detection, the method uses multi-channel pyramids. The computational burden of constructing full pyramids is cleverly avoided by approximating in-between scales from interpolations of the coarser scales. Finally, non-maximum suppression is applied to avoid multiple detections (only a few pixels apart) that correspond to the same person (see Figure 2.1a).

Recently, Deep Neural Networks (DNN) which are potent machine learning tools for pattern recognition inspired by neuronal network models in the brain, were developed to autonomously generate visual characteristic hierarchies. These can implicitly learn highly non-linear and non-convex functions, in an end-to-end manner, and hierarchical feature representations, optimized by training with large annotated datasets for recognizing complex patterns, circumventing the need of explicit feature engineering and selection. Deep learning techniques have been successful in different challenging visual tasks, not only on object detection [171, 124] (see Figure 2.1b), but also on segmentation [78] and tracking [82, 141], having recently surpassed humans in some classification tasks [79] (please, see Appendix A for technical details).

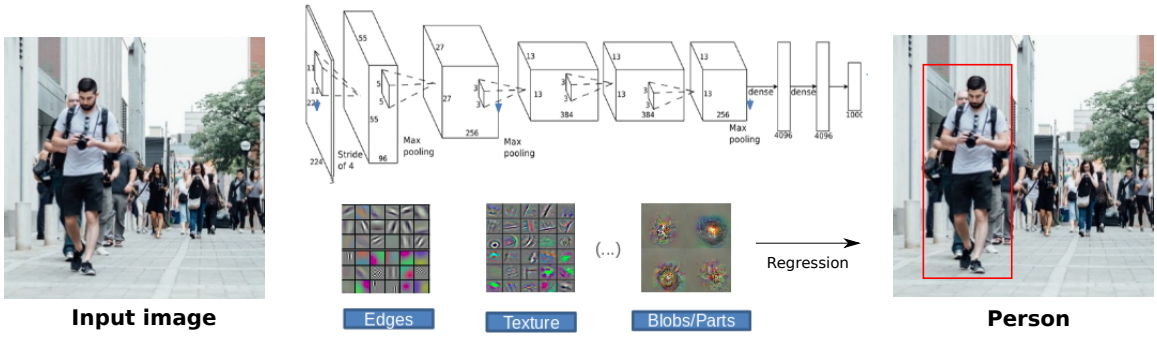
2.2.1 Object Detection with CNNs

The aforementioned network architectures show the progress in object classification tasks. However, we have not yet addressed intuitively more challenging problems such as object detection.

Their proposed method entitled R-CNN [127] first extracts region proposals from the image, and then feeds each region to a CNN with a similar architecture to that of AlexNet [116]. The output of the CNN is then evaluated by a Support Vector Machine (SVM) classifier. Finally, the bounding boxes are tightened by resorting to a linear regression model. This network produces the set of bounding boxes surrounding the objects of interest and the



(a) Classical Approaches



(b) Deep Neural Networks

Figure 2.1: Machine Learning Frameworks for Object Detection

respective classification. The region proposals are obtained through selective search [209]. This method has a major pitfall – it is very slow. This is due to requiring the training of three different models simultaneously, namely the CNN to generate image features, the SVM classifier and the regression model to tighten the bounding boxes. Moreover, each region proposal requires a forward pass of the neural network.

In 2015, the original author proposed Fast R-CNN [70] to address the above-mentioned issues. This network has drastically faster performance and achieves higher detection quality. This is mainly due to two improvements: the first leverages the fact that there is generally an overlap between proposed interest regions, for a given input image. Thus, during the forward pass of the CNN it is possible to reduce the computational effort substantially by using Region of Interest (RoI) Pooling (RoIPool). The high-level idea is to have several regions of interest sharing a single forward pass of the network. Specifically, for each region proposal, we keep a section of the corresponding feature map and scale it to a pre-defined size, with a max pool operation. Essentially this allows us to obtain fixed-size feature maps for variable-size input rectangular sections. Thus, if an image section includes several region proposals we can execute the forward pass of the network using a single feature map, which dramatically speeds up training times. The second major improvement consists of integrating the three previously separated models into a single network. A Softmax layer replaces the SVM classifier altogether and the bounding box coordinates are calculated in parallel by a dedicated linear regression layer.

The progress of Fast R-CNN exposed the region proposal procedure as the bottleneck of the object detection pipeline. A Region Proposal Network (RPN) is a fully convolutional neural network (i.e. every layer is convolutional) [172] for simultaneously predicting objects' bounding boxes as well as objectness score. The latter term refers to a metric for evaluating the likelihood in the presence of an object of any class in a given image window. Since the calculation of region proposals depends on features of the image computed during the forward pass of the CNN, the authors merge RPN with Fast R-CNN into a single network, which was named Faster R-CNN. This further optimises runtime while achieving state of the art performance in the PASCAL VOC 2007, 2012 and Microsoft's COCO [123] datasets. However, the method is still too computationally intensive to be used in real-time applications, running at roughly 7 frames per second (FPS) in a high-end graphics card.

Our work is inspired by [34] which proposed to capture visual attention through feedback DCNN. Similarly

in spirit, we propose a biologically inspired hybrid attention model, that combines bottom-up and top-down mechanisms and, additionally uses artificial human-like foveal vision, to efficiently locate and recognize objects in foveal digital images. More specifically, our method is constituted by three steps: first, we perform a feed-forward pass to obtain the predicted class labels. Second, a backward pass is made to create a saliency map that is used to obtain object location proposals after applying a segmentation mask. Finally, a second feed-forward pass is executed to re-classify the image with selective attention. With a non-uniform foveal visual sensor, the attention is directed to the proposed locations using a foveal spotlight model, whereas for the uniform sensor, the attentional spotlight is oriented in a covert manner to crop patches of the original image.

2.3 Methodologies

Our hybrid detection and localization methodology is depicted in Figure 2.3 and can be briefly outlined as follows: in a first feed-forward pass, a set of object class proposals is computed (Section 2.3.2) and further analyzed via top-down backward propagation to obtain proposals regarding the location of the object in the scene (Section 2.3.2).

More specifically, for a given input image I, we begin by computing a set of object class proposals by performing a feed-forward pass. The probability scores for each class label (1000 in total) are collected by accessing the network's output *softmax* layer. Then, retaining our attention on the five highest predicted class labels, we compute the saliency map for each one of the predicted classes (see Figure 2.4). Then, a top-down back-propagation pass is performed to compute the score derivative of the specific class c. The computed gradient indicates which pixels are more relevant for the class score [188]. In the remainder of this section, we describe in detail the components of the proposed attentional framework.

2.3.1 Artificial Foveal Vision

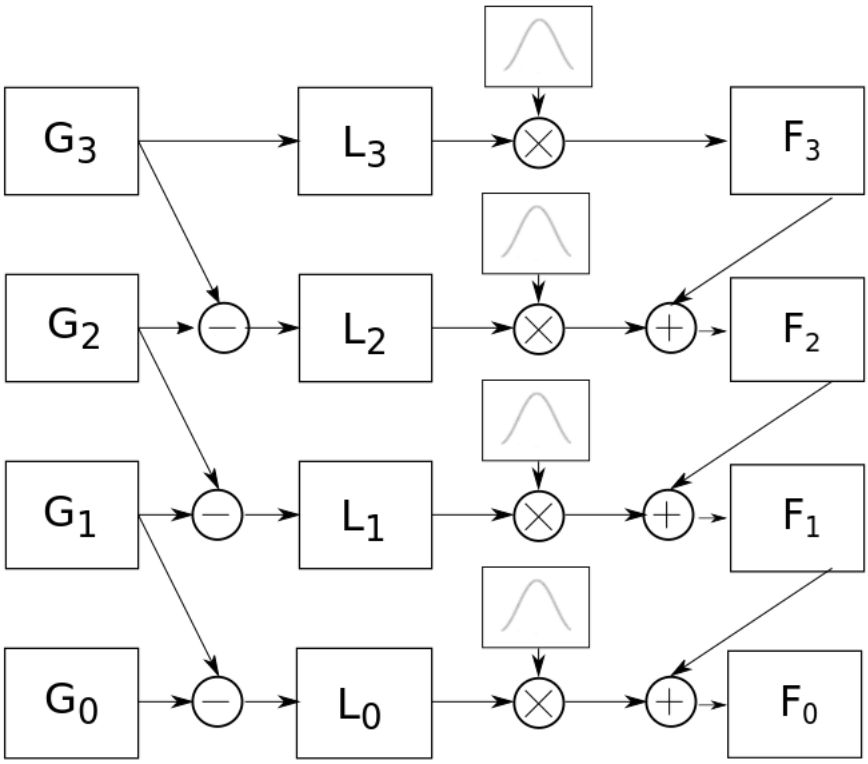
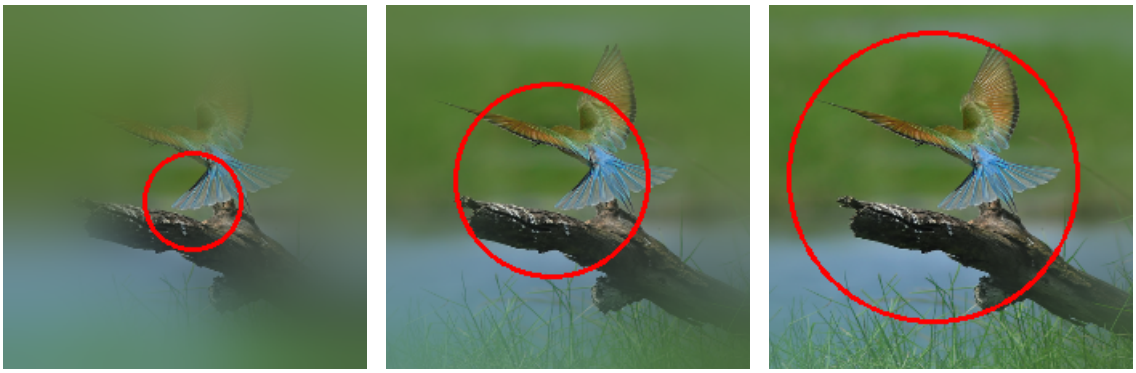


Figure 2.2: A summary of the steps in the foveation system with four levels. The image G_0 corresponds to the original image and F_0 to the foveated image.



(a) $f_0 = 30$ (b) $f_0 = 60$ (c) $f_0 = 90$

Figure 2.3: Example images obtained with our foveation system where $f_k = 2^k f_0$ defines the size of the region with highest acuity (the fovea), from a 227×227 uniform resolution image.

Our foveation system is based on the method proposed in [31] for image compression (e.g. in encoding/decoding applications) which, unlike the methods based on log-polar transformations, is extremely fast and easy to implement, with demonstrated applicability in real-time image processing and pattern recognition tasks [16].

Our approach comprises four steps that go as follows: the first step consists on building a Gaussian scale-space [1] with increasing levels of blur, but similar resolution. The first level (level 0) contains the original image G_0 which is then low-pass filtered with a Gaussian filter g_1 (see 2.2 for the general form of this filter), yielding the image G_1 at level 1.

More specifically, the image G_k can be obtained from the image G_0 via convolution with 2D isotropic and separable Gaussian filter kernels of the form

$$g_k(u,v) = \frac{1}{2\pi\sigma_k^2} e^{-\frac{u^2+v^2}{2\sigma_k^2}} \quad \text{with } 0 \leq k \leq K \tag{2.1}$$

where $\sigma_k = 2^{k-1}\sigma_1$ for $k \geq 1$, and σ_0 is a small value ($\epsilon << 1$) so that $G_0 \approx I$ The Fourier transform of the Gaussian filter kernels is given by

$$\tilde{g}_k(e^{jw_u}, e^{jw_v}) = e^{-\frac{\sigma_k^2}{2}(w_u^2+w_v^2)} \quad \text{with } 0 \leq k \leq K \tag{2.2}$$

where the w_u and w_v are, respectively, the horizontal and vertical spatial frequencies. Note that $\tilde{g}_0 \approx 1$.

Next, we compute a Laplacian scale-space from the difference between adjacent Gaussian levels. The Laplacian scale-space comprises a set of error images where each level represents the difference between two levels of the Gaussian scale-space (see Figure 2.2). Finally, exponential weighting kernels are multiplied by each level of the Laplacian scale-space to emulate a smooth fovea. The exponential kernels are given by

$$H_k(u,v) = e^{-\frac{(u-u_0)^2+(v-v_0)^2}{2f_k^2}}, \quad 0 \leq k \leq K \tag{2.3}$$

where f_0 represents the size of the kernel in the the level 0 of the scale-space, and $f_k = 2^k f_0$ denotes the exponential kernel standard deviation at the k-th level. These kernels are centered at a given fixation point (u_0, v_0) that defines the focus of attention. Throughout the rest of this analysis, without loss of generality, we assume that $u_0 = v_0 = 0$. Figure 2.2 exemplifies the proposed foveation model with four levels and Figure 2.3 depicts examples of resulting foveated images.

[1] in the actual implementation we use a Gaussian pyramid that includes subsampling, but for the sake of the analysis we ignore that step, since the subsampling is done according to Nyquist Sampling Theorem and, thus, has no significant influence in the information content of the images.

Information Reduction

The proposed foveal visual system is a result of a combination of low-pass Gaussian filtering and exponential spatial weighting. To be possible to establish a relationship between signal information compression and task performance, one must understand how the proposed foveation system reduces image information depending on the method's parameters (i.e. fovea and image size).

Low-pass Gaussian Filtering Let us define the original high-resolution image as $I(u,v)$, with size[2] $(N+1)\times(N+1)$ to which corresponds the discrete time Fourier Transform $\tilde{I}(e^{jw_u},e^{jw_v})$. The Fourier transform $\tilde{G}_k(e^{jw_u},e^{jw_v})$ of the filtered image $G_k(u,v)$, at each level k is given by the convolution theorem as follows

$$\tilde{G}_k(e^{jw_u},e^{jw_v}) = \tilde{I}(e^{jw_u},e^{jw_v})\tilde{g}_k(e^{jw_u},e^{jw_v}) \tag{2.4}$$

Following the Parseval's theorem that describes the energy of a signal, the signal energy of the original image $I(u,v)$ is given by

$$E_I = \sum_{u=-\frac{N}{2}}^{+\frac{N}{2}}\sum_{v=-\frac{N}{2}}^{+\frac{N}{2}} |I(u,v)|^2 = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}|\tilde{I}(e^{jw_u},e^{jw_v})|^2 dw_u dw_v. \tag{2.5}$$

The Laplacian at each image level k is given by

$$L_k \ = G_k \ - G_{k+1} \tag{2.6}$$

The Fourier transform of L_k is given by

$$\begin{aligned}\tilde{L}_k \ &= \tilde{G}_k \ - \tilde{G}_{k+1} \\ &= \tilde{I}(e^{jw_u},e^{jw_v})\left(\tilde{g}_k(e^{jw_u},e^{jw_v}) - \tilde{g}_{k+1}(e^{jw_u},e^{jw_v})\right)\end{aligned} \tag{2.7}$$

According to Parseval's theorem, and since $\tilde{g}_k$ and $\tilde{g}_{k+1}$ the energy of L_k is given by

$$E_{L_k} = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}|\tilde{I}(e^{jw_u},e^{jw_v})|^2|\tilde{g}_k(e^{jw_u},e^{jw_v}) - \tilde{g}_{k+1}(e^{jw_u},e^{jw_v})|^2 dw_u dw_v \quad \text{with } 0<k\leq K \tag{2.8}$$

Assuming that $\tilde{I}(e^{jw_u},e^{jw_v})$ has energy equally distributed across all frequencies of the spectrum with magnitude M^2, where M is the amplitude of $\tilde{I}(e^{jw_u},e^{jw_v})$, in the extreme case, has a flat spectrum of magnitude M, the energy E_{L_k} can be expressed as

$$\begin{aligned}E_{L_k} &= \frac{M^2}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\tilde{g}_k^2(e^{jw_u},e^{jw_v}) + \tilde{g}_{k+1}^2(e^{jw_u},e^{jw_v}) - 2\tilde{g}_k(e^{jw_u},e^{jw_v})\tilde{g}_{k+1}(e^{jw_u},e^{jw_v})dw_u dw_v \\ &= \frac{M^2}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{-\sigma_k^2(w_u^2+w_v^2)} + e^{-\sigma_{k+1}^2(w_u^2+w_v^2)} - 2e^{-\frac{\sigma_k^2}{2}(w_u^2+w_v^2)}e^{-\frac{\sigma_{k+1}^2}{2}(w_u^2+w_v^2)}dw_u dw_v \\ &= \frac{M^2}{4\pi^2}\left(\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{-\sigma_k^2(w_u^2+w_v^2)}dw_u dw_v + \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{-\sigma_{k+1}^2(w_u^2+w_v^2)}dw_u dw_v - 2\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{-\frac{w_u^2}{2}(\sigma_k^2+\sigma_{k+1}^2)-\frac{w_v^2}{2}(\sigma_k^2+\sigma_{k+1}^2)}dw_u dw_v\right) \\ &= \frac{M^2}{\pi^2}\left(\int_0^{\pi} e^{-\sigma_k^2 t^2}dt\int_0^{\pi} e^{-\sigma_k^2 t^2}dt + \int_0^{\pi} e^{-\sigma_{k+1}^2 t^2}dt\int_0^{\pi} e^{-\sigma_{k+1}^2 t^2}dt - 2\int_0^{\pi} e^{-\frac{t^2}{2}\left(\sigma_k^2+\sigma_{k+1}^2\right)}dt\int_0^{\pi} e^{-\frac{t^2}{2}\left(\sigma_k^2+\sigma_{k+1}^2\right)}dt\right)\end{aligned} \tag{2.9}$$

[2]without loss of generality let us assume N even

Let us consider the following change of variables $\tau_k = t\sigma_k$, $\tau_{k+1} = t\sigma_{k+1}$ and $\tau = t\frac{\sqrt{\sigma_k^2+\sigma_{k+1}^2}}{2}$. Knowing that the Gaussion error function is given by

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt \tag{2.10}$$

Equation (2.9) becomes

$$\begin{aligned}
E_{L_k} &= \frac{M^2}{\pi^2} \left(\int_0^\pi e^{-\sigma_k^2 t^2} dt \int_0^\pi e^{-\sigma_k^2 t^2} dt + \int_0^\pi e^{-\sigma_{k+1}^2 t^2} dt \int_0^\pi e^{-\sigma_{k+1}^2 t^2} dt - 2 \int_0^\pi e^{-\frac{t^2}{2}\left(\sigma_k^2+\sigma_{k+1}^2\right)} dt \int_0^\pi e^{-\frac{t^2}{2}\left(\sigma_k^2+\sigma_{k+1}^2\right)} dt \right) \\
&= \frac{M^2}{\pi^2} \left(\left(\int_0^{\sigma_k \pi} e^{-\tau_k^2} d\tau_k \frac{1}{\sigma_k} \right)^2 + \left(\int_0^{\sigma_{k+1}\pi} e^{-\tau_{k+1}^2} d\tau_{k+1} \frac{1}{\sigma_{k+1}} \right)^2 - 2 \left(\int_0^{\frac{\sqrt{\sigma_k^2+\sigma_{k+1}^2}}{2}\pi} e^{-\tau^2} d\tau \frac{2}{\sqrt{\sigma_k^2+\sigma_{k+1}^2}} \right)^2 \right) \\
&= \frac{M^2}{\pi^2} \left(\left(\frac{\sqrt{\pi}}{2} \text{erf}\left(\pi\sigma_k\right) \frac{1}{\sigma_k} \right)^2 + \left(\frac{\sqrt{\pi}}{2} \text{erf}\left(\pi\sigma_{k+1}\right) \frac{1}{\sigma_{k+1}} \right)^2 - 2 \left(\frac{\sqrt{\pi}}{2} \text{erf}\left(\frac{\pi}{2}\sqrt{\sigma_k^2+\sigma_{k+1}^2}\right) \frac{2}{\sqrt{\sigma_k^2+\sigma_{k+1}^2}} \right)^2 \right) \\
&= M^2 \left(\left(\frac{\text{erf}\left(\pi\sigma_k\right)}{2\sqrt{\pi}\sigma_k} \right)^2 + \left(\frac{\text{erf}\left(\pi\sigma_{k+1}\right)}{2\sqrt{\pi}\sigma_{k+1}} \right)^2 - 2 \left(\frac{\text{erf}\left(\frac{\pi}{2}\sqrt{\sigma_k^2+\sigma_{k+1}^2}\right)}{2\sqrt{\pi}\sqrt{\sigma_k^2+\sigma_{k+1}^2}} \right)^2 \right)
\end{aligned} \tag{2.11}$$

Since L_k is space invariant and stationary and the total energy can be equaly divided by all pixels, the expected value of the energy per pixel (power), is given by

$$P_{L_k} = \frac{E_{L_k}}{N^2} = \frac{1}{N^2} \sum_u \sum_v L_k^2(u,v) \tag{2.12}$$

Then, by scaling each pixel value by the corresponding exponential kernel, we obtain the foveal image $F_k(u,v)$, at each level

$$F_k(u,v) = L_k(u,v) H_k(u,v) \tag{2.13}$$

Assuming that the Laplacian levels L_k are uncorrelated, the energy E_{F_k} can be approximated by

$$E_{F_k} = \sum_u \sum_v |F_k(u,v)|^2 = \sum_u \sum_v P_{L_k} H_k^2(u,v) \tag{2.14}$$

and the total energy of F becomes

$$E_F = \sum_k E_{F_k} \tag{2.15}$$

and the power of F is given by

$$P_F = \frac{E_F}{N^2} \tag{2.16}$$

Assuming binary coding of pixel values, and according to Shannon-Hartley theorem [186], the maximum transmission rate (channel capacity) of signal is given by

$$C = W \log_2 \left(1 + \frac{P}{Q} \right) \tag{2.17}$$

where W is the bandwidth of the signal, P is the power of the signal, and Q is the power of the quantization noise. In our case, W is equal to $\frac{1}{2}$ cycles per pixel, so that the original image is not aliased. The power of quantization noise, assuming 1 bit quantization with truncation is $\frac{1}{3}$, then the information of F can be approximated by

$$Z_F = \frac{1}{2} \log_2 \left(1 + 3P_F \right) \quad \text{[bits]} \tag{2.18}$$

2.3.2 Weakly Supervised Object Localization

In this subsection we describe in detail our top-down object localization via feedback saliency extraction.

Image-Specific Class Saliency Extraction

As opposed to Itti's [103] that processes the image with different filters to generate specific feature maps, Cao [34] proposed a way to compute a saliency map, in a top-down manner, given an image *I* and a class c. The class score of an object class c in an image I, $S_c(I)$, is the output of the neural network for class c. An approximation of the neural network class score with the first-order Taylor expansion [34][188] in the neighborhood of *I* can be done as follows

$$S_c(I) \geq G_c^\top I + b \tag{2.19}$$

where b is the bias of the model and G_c the gradient of S_c with respect to *I*:

$$G_c = \frac{\partial S_c}{\partial I}. \tag{2.20}$$

Accordingly, the saliency map is computed for a class c by calculating the score derivative of that specific class employing a back-propagation pass. In order to get the saliency value for each pixel (u, v) and since the images used are multi-channel (RGB - three color channels), we rearrange the elements of the vector G_c by taking the maximum magnitude of it over all color channels. This method for saliency map computation is extremely simple and fast since only a back propagation pass is necessary. Simonyan *et al.* [188] shows that the magnitude of the gradient G_c expresses which pixels contribute more to the class score. Consequently, one should expect that these pixels provide us with the localization of the object pertaining to that class, in the image.

Bounding Box Object Localization

Considering Simonyan's findings [188], the class saliency maps hold the object localization of the correspondent class in a given image. Surprisingly and despite being trained on image labels only, the saliency maps can be used on localization tasks. Our object localization method based on saliency maps goes as follow. Given an image *I* and the corresponding class saliency map M_c, a segmentation mask is computed by selecting the pixels with saliency higher than a certain threshold, th, and set the rest of the pixels to zero.
Considering the stain of points resulting from the segmentation mask, for a given threshold, we are able to define a bounding box covering all the non-zero saliency pixels, obtaining a guess of the localization of the object (see Figure 2.4).

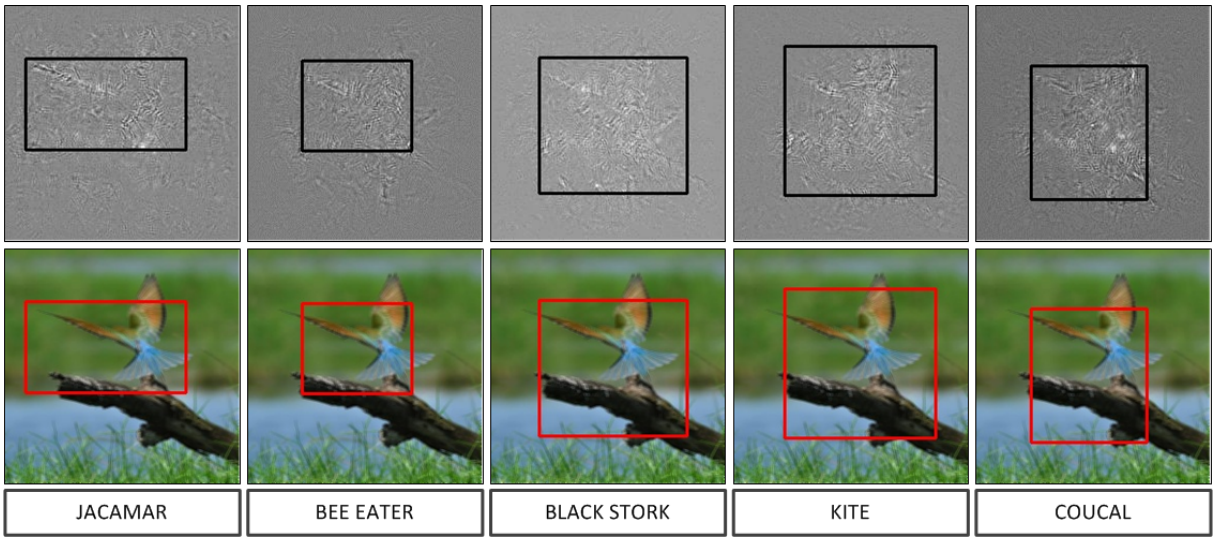


Figure 2.4: Representation of the saliency map and the corresponding bounding box for each of the top-5 predicted class labels of a *bee eater* image of the ILSVRC 2012 data set. The rectangles represent the bounding boxes that cover all non-zero saliency pixels resultant from a segmentation mask with $th = 0.75$.

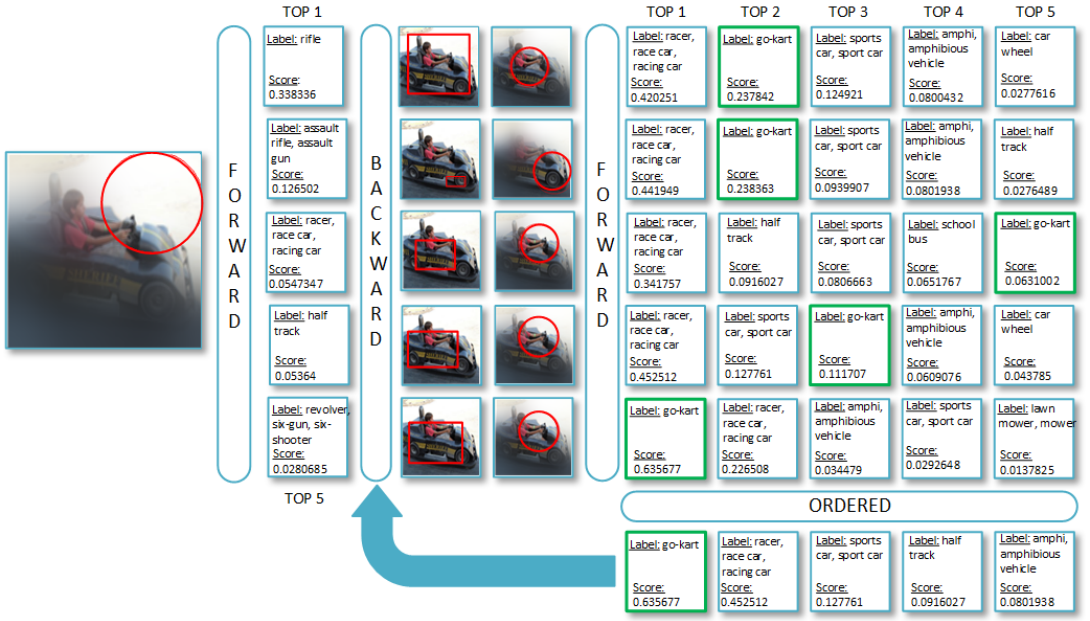


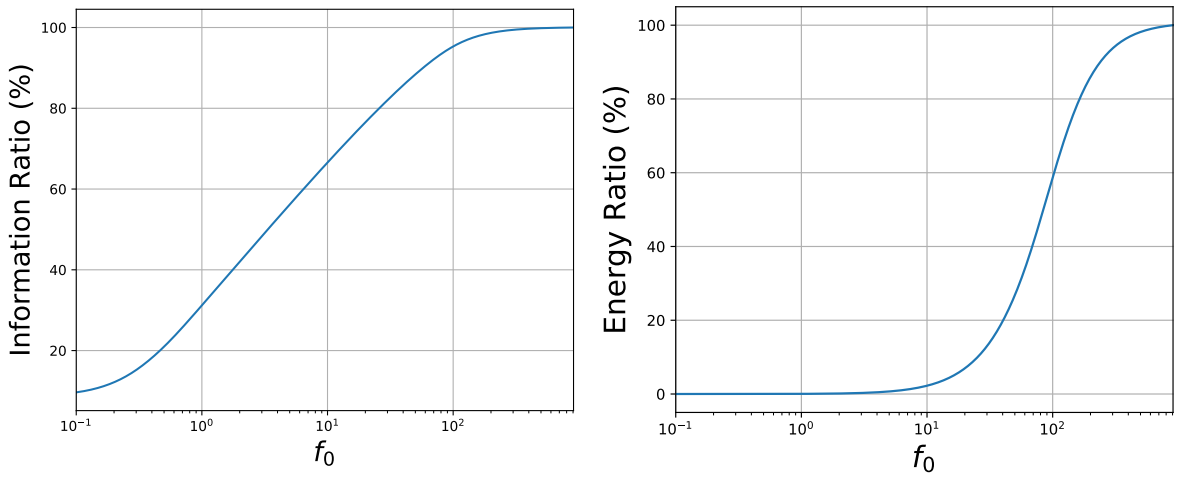
Figure 2.5: Schematic of our iterative refinement model of object detection. First a foveated resized image is loaded into the network to predict the top-5 class labels through a feed-forward pass. Then for each class label, we compute each bounding box with a top-down back-propagation according to the selected threshold. We apply a second foveation centered in each bounding box found and predict again the top 5 class labels with a feed-forward. Given this 25 labels with confidences associated we sort them in descending order, not choosing repeated labels and pick as final solution the top-5. Iteratively we do a re-localization according to those labels with a feedback pass. In our work we only considered two iterations. The red rectangles represent the bounding boxes that contain all pixels above the specified threshold, in this case the threshold was 0.75. The red circles represent the focused area simulating the fovea, that was set to $f_0 = 60$ in this case. The ground truth label of the input image is go-kart.

Given the center of the computed bounding box, we foveate again the original image in the center of the bounding box and iteratively repeat the previous steps (see Figure 2.5).

2.4 Experiments and Results

In this section, we begin by numerically quantifying the proposed non-uniform foveation mechanism information compression dependence on the fovea size. Then, we quantitatively assess the classification and localization performance obtained for the proposed feed-forward and feed-backward passes for various state-of-the-art CNN architectures (section 2.4.2).

2.4.1 Information Compression



(a) Information ratio

(b) Energy ratio

Figure 2.6: Information and energy ratios in function of f_0 for the proposed non-uniform foveal vision mechanism.

In order to quantitatively assess the performance of our methodology, it is important to first quantify the amount of information preserved by the proposed non-uniform foveation mechanism to further understand the fovea size influence in task performance. Through a formal mathematical analysis of the information compression (see section 2.3.1) we can represent the relationship between fovea size (f_0), image size (N) and information compression. In our experiments σ_1 was set to 1, the original image resolution was set to $N \times N = 227 \times 227$ (the size of the considered CNN input layers) and the size of the fovea was varied in the interval $f_0 = [0.1; 227]$. As depicted in Figure 2.6, the information and gain ratios grow monotonically and exhibit a logarithmic behaviour for $f_0 \in [1; 100]$. Beyond $f_0 \approx 100$, the compression becomes residual (information ration of 94%), saturating at around $f_0 \approx 120$. Hence, from this point our foveation mechanism becomes unnecessary since resulting images contain practically the same information as the original uniform-resolution ones.

2.4.2 Attentional Framework Evaluation

In this chapter, our main goal was to develop a single CNN capable of performing, recognition and localization tasks, taking into account both bottom-up and top-down mechanisms of selective visual attention and non-uniform foveal vision. In order to quantitatively assess the performance of the proposed framework we used the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 data set, which comprises a total of $50K$ test images with objects conveniently located in the images center. [3]

Furthermore, we tested the performance of our methods with different pre-trained CNNs (ConvNet) models which are publicly and readily available at Caffe Model Zoo [106], namely, CaffeNet [117], GoogLeNet [197]

[3]source: http://image-net.org/challenges/LSVRC/2012/ [as seen on June, 2017]

and VGGNet [189]. As mentioned in Section 2.3.2, a feed-forward pass is executed originating a vector with the probability distribution of the class label scores. These class labels are used to compute the classification error which compares the ground truth class label provided in ILSVRC with the predicted class labels. Usually, two error rates are commonly used: the top-1 and the top-5. The former serves to verify if the predicted class label with the highest score is equal to the ground truth label. For the latter, we verify if the ground truth label is in the set of the five highest predicted class labels. For a given image, the object location was considered correct if at least one of the five predicted bounding boxes overlapped over 50% with the ground truth bounding box. This evaluation criterion [179] consists on the intersection over union (IoU) between the computed and the ground truth bounding box.

Classification Performance

The classification performance for the various CNN architectures combined with the proposed foveal sensing mechanism are depicted in Figure 2.7a. The CaffeNet pre-trained model which presents the shallower architecture had the worst classification performance. The main reason is that the GoogLeNet and VGG models use smaller convolutional filters and deeper networks enhance the distinction between similar and nearby objects. Regarding the impact of non-uniform foveal vision, a common tendency can be seen for all three pre-trained models. The classification error saturates at approximately $f_0 = 70$. This result is corroborated by the evolution of the information and energy ratio curves, depicted in Figure 2.6, since after 50% ratio the energy ratio slope reduces significantly. This means that on average and for this particular dataset, half of the information contained in uniform resolution images is irrelevant for correct classification. Small size foveas exhibit extremely high error rates, which corresponds to a very small region characterized by having high acuity. This is due to the fact that images that make up the ILSVRC data set contain objects that occupy most of the image area, although the image has a region with high-resolution, it may be small and not suffice to give an idea of the object in the image, which leads to poor classification performance.

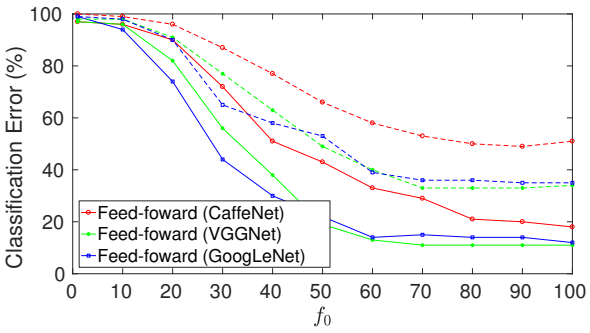
Localization Performance

The localization is considered correct if at least one of the five predicted bounding boxes for an image overlaps over 50% with the ground truth bounding box, otherwise the bounding box is considered wrong. The evaluation metric consists on the intersection over union between the proposed and the ground truth bounding box.

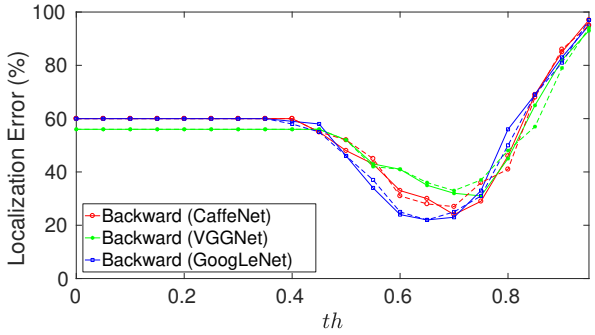
As can be seen in Figure 2.7b, for thresholds smaller than 0.4, the localization error remains consistent and stable at around 60%. From this point, the evolution of the error presents the form of a valley where the best localization results were obtained for thresholds between 0.65 and 0.7. Overall, GoogLeNet presents the best localization performance. We hypothesize that this is mostly due to CaffeNet and VGG models featuring two fully-connected layers of 4096 dimensions that may jeopardize the spatial distinction of image characteristics. Furthermore, GoogLeNet is deeper than the aforementioned models and hence can learn discriminant features at higher levels of abstraction.

Sequential Fixations

Our final goal was to assess if there was a significant gain in performance for sequential fixations. In order to understand how the foveation point of the first feed-forward pass influences the classification error, we made it vary along a 8×8 grid. The threshold applied to the segmentation mask was fixed to $\theta = 0.7$, the size of the fovea varied between 0 and 180 and the classification error was calculated for each position over all considered $f_0 \in \{0, 100\}$. In Figure 2.8 we compare the classification error between first and second feed-forward passes as a function of the fixation point. Since the objects of the data set are mainly centered, the classification error is smaller in the center, as expected. However, from the first to the second pass the accuracy improves, independently of the initial foveation point, demonstrating the advantages of having a system capable of manipulating the fixation point.

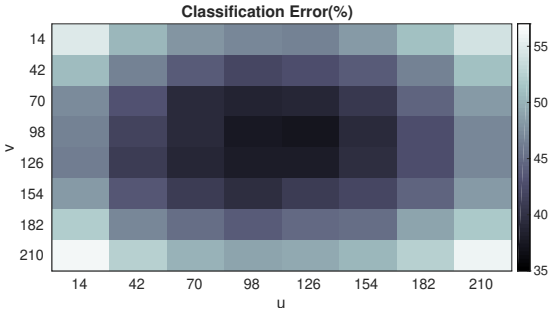


(a) Classification error

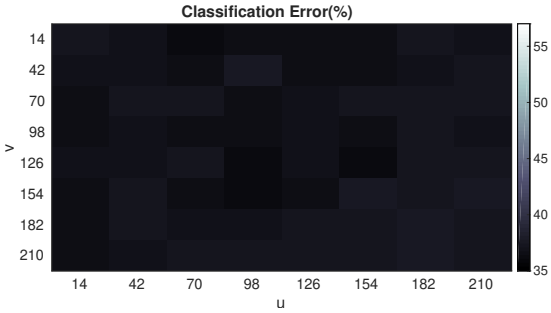


(b) Localization error

Figure 2.7: Classification and localization performance for various network architectures and sensing configurations. Left column: Dashed lines correspond to top-1 error and the solid ones correspond to top-5 error. Righ column: Dashed lines correspond to $f_0 = 80$ and solid lines to $f_0 = 100$.



(a) First pass



(b) Second pass

Figure 2.8: Classification performance in function of initial foveation point (u_0, v_0) where dark and bright represent better and worse performance. The classification error was calculated over all f_0 and fixing $\theta = 0.7$

In order to better understand how the foveation size affects the classification error both for centered and non-centered foveation points we fixed $\theta = 0.7$ and varied f_0 between 0 and 180. In Figure 2.9 we verify that the accuracy improvement between the first and the second feed-forward classification is not significant when the foveation is centered, corresponding to a maximum of 10%. However, when the foveation is non-centered (average of all foveation positions) the maximum accuracy gain between the two passes is 43%.

To understand the effect of the threshold applied on the segmentation mask, on the localization performance, we fixed the fovea size to $f_0 = 70$ and varied the threshold in the interval $\theta \in \{0, \ldots, 1.0\}$. As observed in Figure 2.10(a) there is neither gain between backward passes nor differences between foveating in the center or elsewhere in the image. Localization tasks depend mostly on the low frequency of the image signal, thus, when we foveate an image we only remove high frequencies outside the fovea, however the location of the object remains detectable. For mask thresholds smaller than 0.4, the localization error remains stable. From this point, the evolution of the error presents the form of a valley obtaining the lowest localization error for thresholds of 0.65 and 0.7. This shows that exists a sensitive trade-off for the threshold selection, for accurate bounding box selection. For this reason, we chose $\theta = 0.7$ to lead to minimum errors, when varying the fovea size as illustrated in Figure 2.10(b).

2.5 Conclusions

In this work we proposed a biologically inspired framework for object classification and localization that combines human-like artificial foveal vision with bottom-up and top-down attentional mechanisms, using DCNN architectures. Through the analysis performed in our tests, we conclude that deeper neural networks present better performance when it comes to classification.

The main experimental goals of this study were: first, to assess the performance impact of mimicking non-

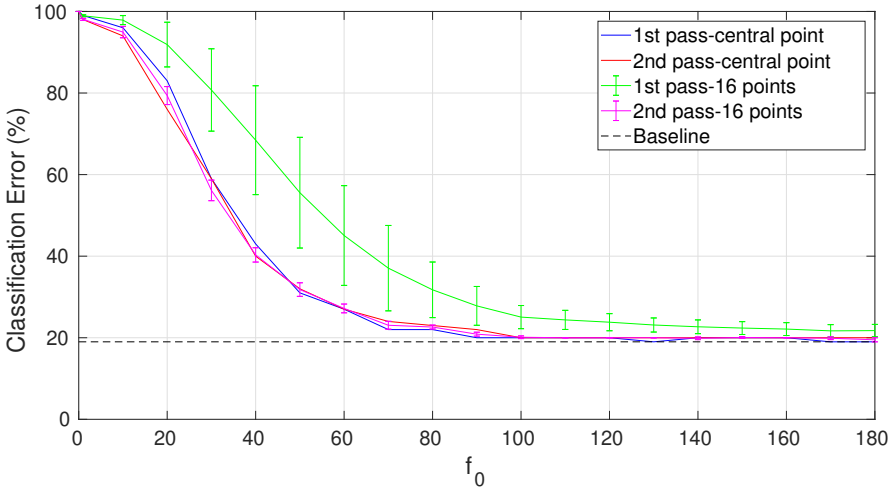
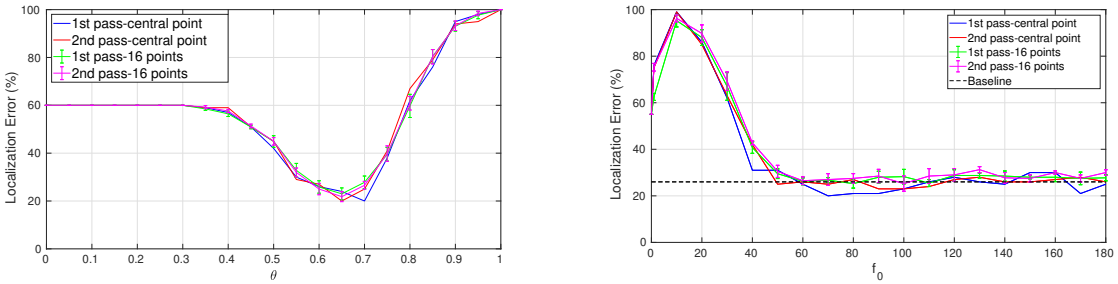


Figure 2.9: Classification Performance in function of the fovea size f_0 with $\theta = 0.7$. The baseline was computed with $f_0 = 227$ (the resolution of the input image) to simulate an input image without any blur corresponding to minimum error



(a) (b)

Figure 2.10: Localization Performance: (a) in function of the threshold applied to the segmentation mask with a fixed fovea size $f_0 = 70$; (b) in function of the fovea size f_0 where the threshold applied to the segmentation mask was set to $\theta = 0.7$ since it results in a minimum localization error.

uniform, human like, foveal vision mechanisms in recognition and localization tasks, when combined with state-of-the-art CNN architectures. Second, to verify if a simple saliency-based fixation point control mechanism would improve classification performance across fixations. We concluded that from a certain information compression level, proportional to the fovea size, the performance in classification task saturates and that sequential fixations using a saliency based saccadic mechanism improves task performance.

Furthermore, the results obtained for non-uniform foveal vision are promising. From a given fovea size (f_0), the performance in classification tasks saturates. On the one hand, when using a methodology that replicates human visual behavior, it is necessary to use successive foveations (saccades). This is because in real scenarios, objects can be located anywhere in the image, and the results show that the classification performance improves significantly from the first to the second feed-forward passes while localization does not. Location only depends on lower image frequencies and smoothing them with our foveation mechanism does not significantly affect performance. We emphasize that the goal of this work was studying the impact of information reduction via space-variant blurring of the original image, on classification and localization tasks using a state-of-the-art CNN classifier.

Also, the proposed approach does not directly affect computational performance, since the number of pixels in the input images is fixed. Alternatively, one could leverage recently proposed retina-DCNNs [155] that are capable of learning and inference in cortical domain, using more compact and computationally efficient log-polar representations.

Although the proposed approach does not provide any computational complexity gains, when compared with

methods based on log-polar transforms [155] we demonstrated that filtering-based foveation mechanisms have similar performance to conventional uniform resolution methods, when combined with saccadic mechanisms, and thus may be beneficial for image transmission in low-bandwidth communication channels, by reducing the amount of transmitted data, when on-board computational resources are low..

Finally, the current system is memoryless and performance would benefit of integrating task-related evidence accross saccades using probabilistic fusion approaches.

The main published contributions of our work are twofold:

1. The first published in **P.I**, lies in the evaluation of the performances obtained with different non-uniform blurring. We were able to establish a formal relationship between classification and localization performance, and the different levels of information preserved by each of the sensing configurations, for different CNN architectures. A set of experiments with objects centered in the fovea, demonstrate that it is not necessary to store and transmit all the information present on high-resolution images since, beyond a certain amount of preserved information, the performance in classification and localization task saturates.

2. Second, we demonstrate that when mimicking the human visual foveation mechanism with the proposed model, saccades are necessary to improve both recognition and localization performance, since we demonstrate in **P.II** that for non-centered objects, the gain in classification performance between iterations is significantly improved.

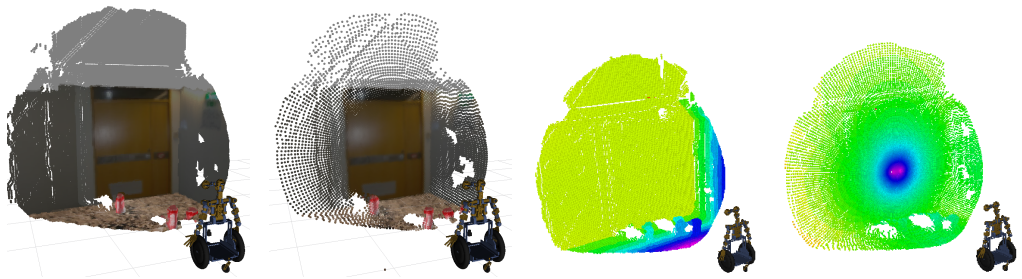
Future work The current limitations of the proposed methodologies are twofold. First, while the input images of the DCNN network are non-uniform foveal images, the network was trained with conventional uniform ones. In the future, we intend to enhance the pipeline by considering the following ideas:

- *Foveal ImageNet*: fine-tuning the network with foveated versions of the ImageNet dataset [118]. One should expect to see large improvements, in particular for close objects whose overall characteristics become unperceivable as the level of detail decays very rapidly towards the periphery.

- *Mimicking human gaze patterns*: our current gaze control mechanism is based on the simple idea of prioritizing ocular attention to task-dependent salient regions of the visual stimuli, without information integration across fixations. Another research line in the literature frames saccadic eye movement modeling as a learning problem, and use either human demonstrated gaze patterns gathered with eye tracking technologies [149] or reinforcement learning techniques to learn task-specific gaze control policies (e.g. finding a specific object) [141]. In the future we intend to learn from human demonstrations, to sequentially fixate image sequences, given the task of finding a pre-specified object.

Chapter 3

3D Visual Search with Foveal Vision and Space-variant Spatial Representations



(a) RGB-D - Cartesian (b) RGB-D - Foveal (c) Uncertainty (d) Uncertainty

Figure 3.1: A snapshot of the RGB-D point clouds and associated probabilistic measures obtained with the proposed Cartesian and foveal stereo sensor models. Blue and purple colors correspond to higher precision measurements.

The goal of this chapter is to study and to develop biologically inspired 3D stereo vision mechanisms for 3D object reconstruction tasks. More specifically, we study how information provided by foveated images sampled according to the log-polar transformation can be integrated over time in order to build accurate world representations and accomplish visual search tasks in an efficient manner. We focus on a specific visual information modality – depth – and on how to store it in a flexible memory structure. We propose a probabilistic observational model for a stereo system that relies on the Unscented Transform in order to propagate uncertainty in stereo matching, due to spatial quantization in the retina, to the 3D Cartesian domain. Probabilistic depth measurements are integrated in a novel Sensory Ego-Sphere whose topology can be biased with foveal-like distributions, according to the autonomous agent short-term tasks and goals. Furthermore, we investigate an Upper Confidence Bound (UCB) algorithm for the task of simultaneously finding the closest object to the observer (visual search) and learning the surrounding environment 3D map (mapping). The performance of task execution is assessed both with a foveated log-polar sensor and a classical uniform one. The advantage of foveal vision and customized ego-sphere representations are illustrated in a series of experiments with a realistic simulator. The idea of using human like stereo vision is novel within the robotics community, and beneficial when combined with saccadic mechanisms, with demonstrated improved 3D reconstruction accuracy.

3.1 Introduction

In this work we propose a probabilistic selective attentional framework for artificial systems provided with binocular foveal vision. Our framework relies on visual information and associated confidence measures (see Figure 3.1) that are used to autonomously drive the agent's gaze direction during search tasks. Our contributions are the following. First, we model the stereo reconstruction uncertainty that arises as a result of spatial quantization phenomena inherent in the retina. Our approach considers Gaussian Receptive Fields[1] (RFs) distributed in space following two different tessellations: (i) a classical uniform (Cartesian) arrangement and (ii) a log-polar one that mimics the human retina. The RFs in the latter present a space-variant spatial distribution and support radius [157]. The Unscented Transform (UT) [107] is used to propagate belief from the 2D retina domain to 3D via stereo reconstruction. When compared with previous approaches that also assume Gaussian quantization noise and that rely on first order linearizations to approximate the non-linear transformations involved in 3D reconstruction [115], our method based on the UT is more precise and hence improves 3D estimation quality. Second, the probabilistic sensory measurements are integrated in a novel versatile randomized SES whose topology can be biased according to the autonomous agent short-term tasks and goals. The proposed SES helps achieving the task, by allocating the limited resources more densely to important surrounding regions according to the task. Finally, a decision-making process, framed within a multi-armed bandit setting [10], acts as a mediating cognitive attentional process that seeks to maximize expected task-related rewards. The proposed decision making algorithm relies on statistical

[1] *Receptive fields* are the fundamental visual processing units. Each corresponds to a specific region in the retina (image) and is represented by the average value of the photo-receptors (pixels) within it (e.g. average color). For more details, we refer the interested reader to [49].

measures to decide where to look next by selecting the most promising regions to attend. We investigate a simple UCB algorithm [3] for the task of finding the closest object to the observer. The UCB algorithm controls the exploration-exploitation trade-off typical of decision under uncertainty algorithms: to accomplish the task it is necessary to explore the world, but too much exploration will delay the task execution.

The remainder of this chapter is organized as follows. In section 3.2 we conduct a brief overview of the attentional frameworks available in the literature with a strong emphasis on probabilistic-based methodologies. In section 3.3.2, we outline the proposed sensor observation model and the uncertainty propagation model from the retinal domain to 3D. In section 3.3.3, we introduce a novel biologically inspired short-term memory structure which is egocentric, compact, and convenient for fast and efficient information update and retrieval. In section 3.3.4, we endow our system with a decision-making process that actively drives the agent's gaze direction, through sequential saccadic eye movements. Finally, in section 3.4, we experimentally validate our model and compare a conventional Cartesian camera against a space-variant vision system. The obtained results demonstrate that a wider field of view at the cost of less peripheral resolution is advantageous in visual-search tasks. We show that with our methodologies different gaze patterns emerge depending on the sensor characteristics and decisions on confidence bounds. Furthermore, we demonstrate that spatial memory biases, reflecting prior knowledge about the world structure and the task at hand, allow large performance improvements in visual search tasks.

3.2 Related Work

Probabilistic based active vision requires not only the characterization of the sensory-motor uncertainties, but also the definition of memory structures that facilitate continuous recall and temporal fusion of probabilistic sensory data. Therefore, we organize the present section in two distinct parts. At first we overview the state-of-the-art in active vision with an emphasis on probabilistic models of overt attention. Afterward, we analyse the memory data structures proposed in the literature suitable for applications related to attention.

3.2.1 Active Vision

It has been shown that visual search tasks are computationally prohibitive due to their combinatorial nature [207] and that the attentional mechanisms are responsible to drive the perceptual search problems tractable by deciding which stimuli to enter the cognitive apparatus through efficient resource allocation. In this work we focus on 3D active sensing which is tightly coupled to the concept of overt attention. One goal of overt active vision mechanisms is to direct the vision apparatus towards locations such that:

- the information about the surrounding environment is increased over time (exploration);
- the desired region is centered in the images of the stereoscopic system (e.g. eyes) and thus observed by the retinal zone of maximum visual acuity (exploitation).

Next-Best-View Planning

One approach to the active vision problem is to sequentially compute the NBV in 3D space, according to some criteria related to task performance (e.g. reduce entropy in 3D reconstruction) and then, move the sensor towards that location. For example, [27] proposes a simple NBV algorithm which greedily targets the gaze of a humanoid robot at points of maximum entropy along a trajectory. In the context of active 3D reconstruction [44], existing NBV approaches belong to one of two main categories: frontier-based and information-driven planning. Frontier-based planners [221, 47] guide the robot to boundaries between unknown and free space, which implicitly promotes exploration. Information-driven methods back-project probabilistic volumetric information on candidate views via ray casting, and select the views that maximize expected information gains [44]. Methods differ in the way they define information gain. The approach leads to interleaved gazing and path planning that converges to a high

confidence free-space robot trajectory plan. The authors of [114] propose to use the average depth information theoretic entropy over all voxels traversed via ray casting. Instead of just considering the entropy, [100] proposes a set of extensions to [114]'s information gain definition, including the incorporation of visibility probability as well as the likelihood of seeing new parts of the object.

More sophisticated NBV planning models are framed within probabilistic paradigms that account both for sensori-motor uncertainties as well as the world intrinsic stochasticity and unpredictability. A common idea behind these models is that statistical objectives are the fundamental driving elements behind visual attention. From a Bayesian standpoint, attention seeks to actively infer the future actions that maximize the expected information gain given the spatio-temporal context. Therefore informational gain is itself the inner goal behind attention [64]. The probabilistic-based saliency model proposed in [101] suggests that surprising events or stimuli attract attention. The *Kullback-Leibler* (KL) divergence between prior and posterior beliefs is by convention used as a measure of surprise. However, surprise models are purely exogenous by nature since they react to observed stimuli. Active vision models based on optimal stochastic control principles pose the action selection problem within *Bayes* risk minimization framework, and differ on the chosen policies. On one hand, infomax algorithms [33] seek to maximize the expected accumulated future informational gain in fixed time-horizon. On the other hand greedy MAP policies consider only a one-step look ahead time window [148] and self-knowledge about the retinal acuity map to decide the best location to attend. A recent work on active sensing accounted also for behavioral costs [4], such as the energy and temporal costs incurred in choosing a given motor action.

Despite the demonstrated applicability of the previously mentioned approaches on target search tasks in monocular images, there are no works studying depth cues inferred by stereo vision, and the influence of foveal vision in the search strategies on binocular setups. The stereo reconstruction problem using foveated images has been addressed in the literature, namely in [20], where the authors have shown that it is possible to compute dense disparity maps from log-polar images. Nevertheless, with foveal images, stereo matching accuracy degrades in the image periphery. This motivates the need for modeling depth uncertainty in stereo reconstruction, due to space-variant discretization in foveated images and use this uncertainty to decide where to look next.

In this work we analise the ability of active foveal stereo systems to accurately map the environment and efficiently execute visual search tasks. Some visual tasks are more naturally represented in 3D, for instance the search for nearby objects, as illustrated in this work. Therefore, the main contribution of this work is the formulation of visual search tasks in 3D and the development of novel methods for uncertainty propagation and spatial representations required for this purpose. We show that adequate retinal topologies and 3D spatial representations play a role in the speed of execution and accuracy on localization of targets in 3D search tasks while keeping the computational resources under control.

3.2.2 Spatial Memory Data Structures

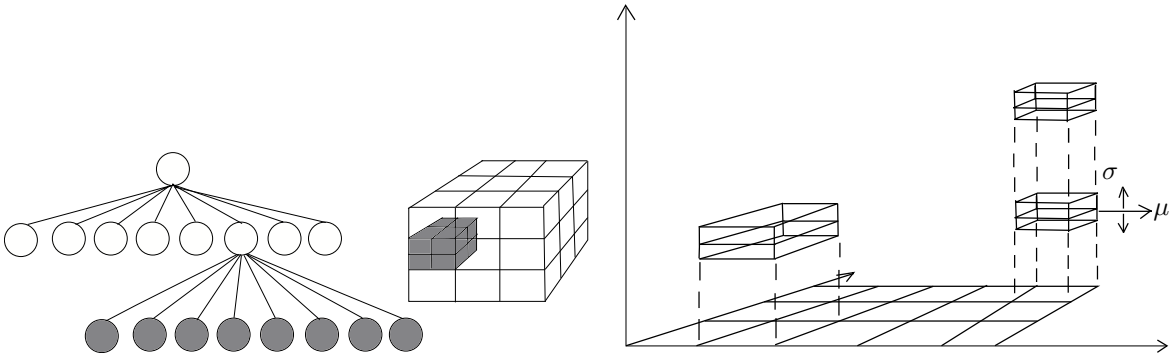
Among many different domains, cognition is focused on abilities to deal with spatial knowledge, namely with relations between objects in space and has been widely studied in psychology and neuroscience [190]. These abilities require remembering and encoding spatial information of everyday object locations, in different reference coordinate frames, depending on the task [38].

In decision-making problems involving perception, autonomous agents rely on spatial memory structures to continuously store and query probabilistic information in a robust and efficient manner. Spatial memory, thus plays a key role and is a core component of any cognitive architecture and may belong to one of two main categories: allocentric or egocentric, depending on the used frames of reference [72].

Allocentric Representations

Allocentric representations specify relative locations between objects in landmark or object centered reference frames, independent of the agent location. The most common allocentric representations for spatial mapping

available in the literature are occupancy grids. Such maps represent the environment as uniform blocks of cells, each cell having a binary state (either occupied, or free). They are popular in the robotics community since they simplify collision checking and path planning, access is fast and memory use can be made efficient through octree-based geometric modeling [133], in particular, the 3D voxel-based OctoMap structure of [90] (see Figure 3.2a). Elevation maps [83] is a compact 2.5D probabilistic representation that encodes continuous heights on 2D grids [136], offering a convenient representation for legged locomotion [76]. However, they are unsuitable for complex environments where the agent may have to navigate between objects at distinct heights (e.g. a ladder, a structure with several levels or floors). To overcome this limitation, multi-level Surface maps [205] have been proposed. These consider a list of surface patches for each grid cell. Still, their main setback resides on the impossibility of modeling free space (see Figure 3.2b). Recently, the idea of using continuous representations in mapping has also attracted great attention from the robotics community [152]. For example, the authors in [151] proposed the use of Gaussian Proccesses (GPs) to encode interdependence between cells and thus correlations between structures in the environment. While continuous mapping techniques based on GPs offer a convenient framework for exploration via Bayesian inference they lack in computational efficiency, since they rely on Bayesian Optimization (BO) techniques [185] in high-dimensional spaces. Recent work of [104] sets on the promising idea of considering fewer observations for close to real-time inference. Still, GPs require intensive sampling during collision checking for motion planners, which could be prohibitive for real-time applications.



(a) OcTree (b) Multi-level Height Maps

Figure 3.2: Allocentric Representations

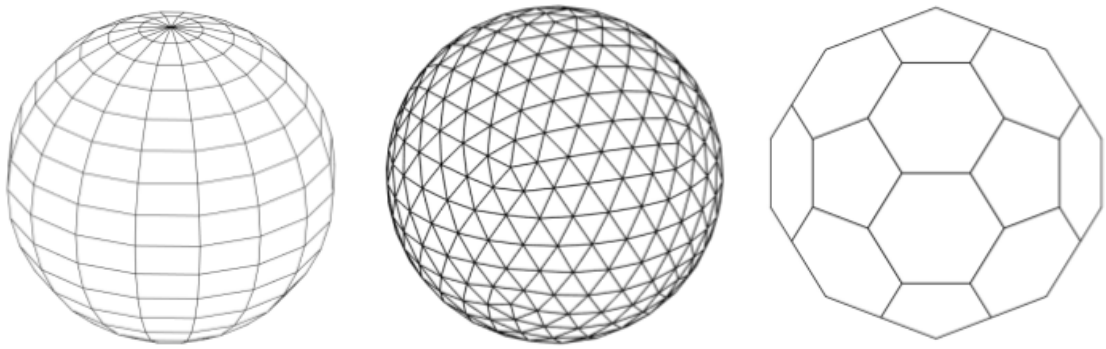
Egocentric Representations

Egocentric or viewer centric reference frames encode locations with respect to the agent body coordinates, and are appropriate for planning and performing motor actions within the peripersonal space, namely for visuomotor reasoning and coordination during reaching and manipulation of objects.

Egocentric representations [161] of space are convenient for cognitive attention modeling and multi-modal sensory data fusion as a short-term memory, and have been extensively used in robotics [59, 178]. From a practical stand point, egocentric spherical representations offer several advantages when compared to typical Cartesian representations such as regular occupancy grids, point clouds or OcTrees [89]. Spherical representations based on ego-centric polar coordinate systems, are typically more compact (low-memory requirements), by projecting the surrounding 3D world on a 2D spherical manifold, and avoid the requirement of computationally expensive ray-casting techniques to deal with visibility issues. However, these advantages come at the cost of expensive updates every time the observing agent moves.

Different representations and data structures for spherical egocentric representations have been proposed in the literature. Typically, 2D array type structures based on spherical coordinate systems are used to represent the spherical surface [178] (see Figure 3.3a). These can be accessed in $\mathcal{O}(1)$ time hence being appropriate for real-time applications. Yet, they are non-isotropic and therefore data is not stored uniformly over the surface

(i.e. the resolution is higher near the poles) (see Figure 3.3b). On the other hand, the geodesic dome type data structure [161] is isotropic and can better approximate 3D shape. However, indexing becomes less trivial and less efficient due to its non-regular topology. To tackle this issue [86] proposed a hierarchical geodesic structure that can significantly speed-up access times. In another work [53] the authors proposed an egocentric log-spherical grid named Bayesian Volumetric map, that was proven suitable for probabilistic multi-modal sensor fusion. Other less uniform spherical polyhedra include icosahedral tessellations of the sphere (see Figure 3.3c). All of the aforementioned forms are highly regular and structured, limiting their flexibility to implement arbitrary tessellations.



(a) 2D Array (b) Geodesic Dome (c) Truncated Icosahedron

Figure 3.3: Egocentric Representations

Nevertheless, none of the previous allocentric and egocentric representations can be easily reconfigurable and is suitable for the incorporation of task-dependent priors to enhance or impair storage and recall of information [41]. This fact motivates the need for developing more sophisticated and versatile spatial memory structures that should facilitate giving more importance to particular regions or orientations, encoded in either allocentric or egocentric frames of reference, depending on the nature of the task. For instance, while crossing a street people prioritize their visual sensorimotor resources to antipodal lateral directions (car detection), while when climbing a stair, people prioritize resources to bottom directions to detect the steps.

Typical tessellations of the sphere include quasi-uniform icosahedral tessellations, less uniform spherical polyhedra or non-uniform latitude/longitude grids. All these forms are highly regular and structured, which limits their flexibility to implement arbitrary shapes. The method proposed in this work is based on projecting in the sphere randomly generated points according to a mixture of 3D Gaussian distributed points with an arbitrary number of components, focal points (means) and dispersions (covariances), representing sampled directions. This generates an irregular grid but we can define more freely areas on the sphere with varying degrees of density and dispersion. Our sampling scheme is easy to implement and allows for the creation of task-biased sensory egospheres. As opposed to previously proposed deterministic counterparts, our SES relies on an easy to implement random sampling scheme that allows for fast creation and real-time access of arbitrary re-configurable topologies.

3.3 Methodologies

In the proposed problem (see Figure 3.4), the observer's goal is to select the oculomotor actions that maximize task related rewards. On one hand we rely on a recursive Bayesian filter that sequentially accumulates sensory inputs and extracts valuable information about the agent and the environment state, given noisy observations. On the other hand, a decision-making algorithm predicts the best future locations to gather information, according to

some statistical or behavioral criteria.

3.3.1 System Overview

The environment structure, i.e. 3D map, is a projection of the world structure $W \subset \mathbb{R}^3$ in the agent's egocentric reference frame $\mathcal{E}$, internally represented by a discrete set of points, each associated to a specific observation direction (see section 2.2). Let us denote the set of environment sample points by

$$X_t = \{\mathbf{x}_t^i \in \mathbb{R}^3, i = 1, ..., N_x\} \tag{3.1}$$

where N_x is the total number of considered observation directions. These points are modeled as Gaussian random variables, initialized with mean and covariance selected according to *a priori* knowledge about the type of environment in which the robot operates. The egocentric reference frame $\mathcal{E}$ is head-centered, has three translational degrees of freedom and fixed orientation with respect to the world frame of reference (see Figure 3.5).

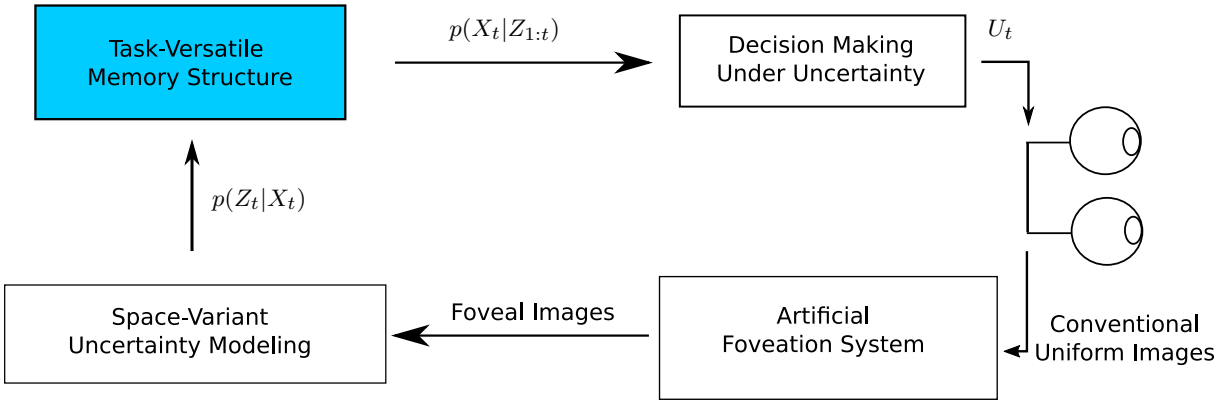


Figure 3.4: General diagram describing the proposed probabilistic binocular active vision framework.

In order to execute visual search tasks, the proposed cognitive architecture is equipped with two sensory-motor modalities:

- proprioception - provided by odometric and oculocephalic joint encoders;
- stereo vision - provided by a stereo camera system.

The observer is allowed to change its state, i.e. the observation view point, through base and oculocephalic movements (see Figure 3.1). At each time instant, the proprioceptive modality reports the robot base location and its internal kinematic state. More specifically, the robot position and orientation $P \in \mathbb{R}^6$ in the inertial frame of reference $\mathcal{W}$, the agent's eyes horizontal vergence ($\theta_t^{\mathrm{v}} \in \mathbb{R}$) and the head pan and tilt joint angles ($\theta_t^{\mathrm{p}}, \theta_t^{\mathrm{t}} \in \mathbb{R}$). Let us denote the joint set of odometric and oculocephalic measured/controlled joint positions by

$$U_t = \{P_t, \theta_t^{\mathrm{v}}, \theta_t^{\mathrm{p}}, \theta_t^{\mathrm{t}}\} \tag{3.2}$$

We assume that the proprioceptive modality provides noise-free observations. In other words, we consider that the measurement errors are negligible with respect to the visual sensor errors and therefore that the robot location and kinematics, and thus, the transformations between the various reference frames involved in our system (see Figure 3.5), can be deterministically determined from U_t. Furthermore, we assume that the environment $\mathcal{W}$ is static for the duration of the search task and is not affected by the robot motor actions U_t (the base location and the posture of the robot's head). The preceding assumptions yield the following probabilistic simplification

$$p(X_t|W, U_t) = p(X_t|{}^{\mathcal{E}}\mathbf{R}_{t,\mathcal{W}} W + {}^{\mathcal{E}}\mathbf{t}_{t,\mathcal{W}}) = p(X_t) \tag{3.3}$$

where ${}^{\mathcal{E}}\mathbf{R}_t, \mathcal{W} \in \mathbb{R}^{3\times3}$ and ${}^{\mathcal{E}}\mathbf{t}_{t,\mathcal{W}} \in \mathbb{R}^{3\times1}$ are an orthogonal rotation matrix and a translation vector, respectively, obtained by combining deterministic proprioceptive joint angle measurements with known forward kinematics. The stereo sensor computes a list of 3D point estimates Z_t defined in a cyclopean reference frame $\mathcal{C}$, with origin at the midpoint of the stereo baseline, from noisy point correspondences observed in the left and right retinal domain. Let us denote the set of 3D points by

$$Z_t = \{\mathbf{z}_t^o \in \mathbb{R}^3, o = 1, ..., N_{v,t}\} \tag{3.4}$$

where $N_{v,t}$ is the total number of observed independent and identically distributed (i.i.d.) measurements by the stereo sensor at time t. The observation model described in section 3.3.2 explains how measurements Z_t are generated according to the environment 3D structure, egocentric projection X_t:

$$Z_t \sim p(Z_t|X_t) \tag{3.5}$$

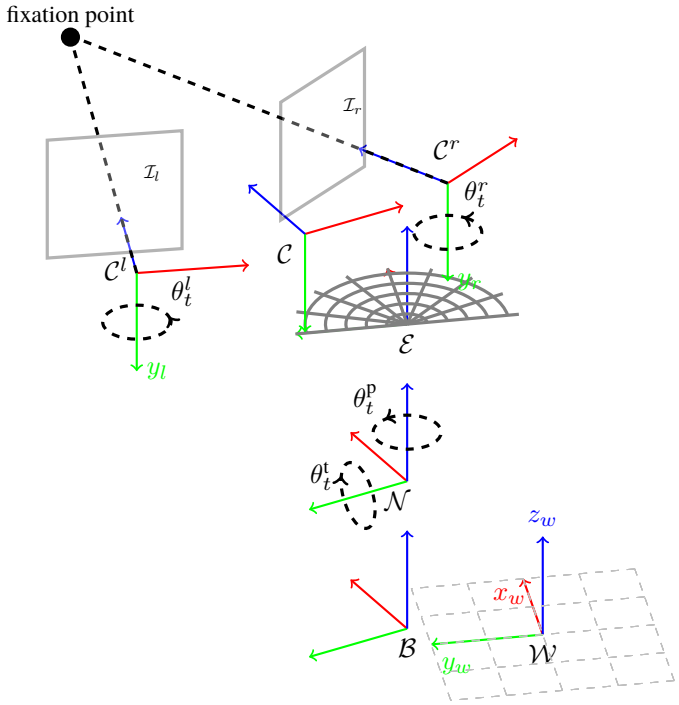


Figure 3.5: The various coordinate systems used by our system (best seen in color): The inertial world frame ($\mathcal{W}$) in which the environment is represented; the base frame ($\mathcal{B}$) which is rigidly attached to the mobile robot base, and permits determining the robot pose in the world, given the odometric readings; the neck frame ($\mathcal{N}$) which allows representing pan and tilt cephalic movements; the egocentric frame ($\mathcal{E}$), which is fixed and defined during initialization time, in which spatial memory is defined and sensor fusion performed; the cyclopean frame ($\mathcal{C}$) in which stereo observations are represented; the convergent, non-parallel pair of camera frames ($\mathcal{C}^l$,$\mathcal{C}^r$), in which monocular images are obtained.

3.3.2 Stereo Sensor Model

In stereo vision, a general stereo matching algorithm computes a set of one-to-one point correspondences between two images [200]. However the precision of the measurements is finite and constrained by the fundamental image-sensing units size and spacing. In order to model reconstruction uncertainty due to the limited sensing precision at the retinal level we consider a probabilistic observation model for our stereoscopic sensor [160].

Nonparallel Stereo System

Let us suppose that our stereoscopic system is composed by a convergent, non-parallel pair of pinhole cameras $\mathcal{C}^l, \mathcal{C}^r$, allowed to rotate around their y optical-axis by $\theta^l = \frac{\theta^v}{2}$ and $\theta^r = -\frac{\theta^v}{2}$, respectively, and are separated by a fixed baseline b. Furthermore, let us assume that the stereo system is calibrated, and thus, the intrinsic $\mathbf{K}^l$, $\mathbf{K}^r$ and extrinsic $\mathbf{R}_y(\theta^v)$, $\mathbf{T}(b)$, camera parameters are always known.

Gaussian Stereoscopic Retinal Observation Model

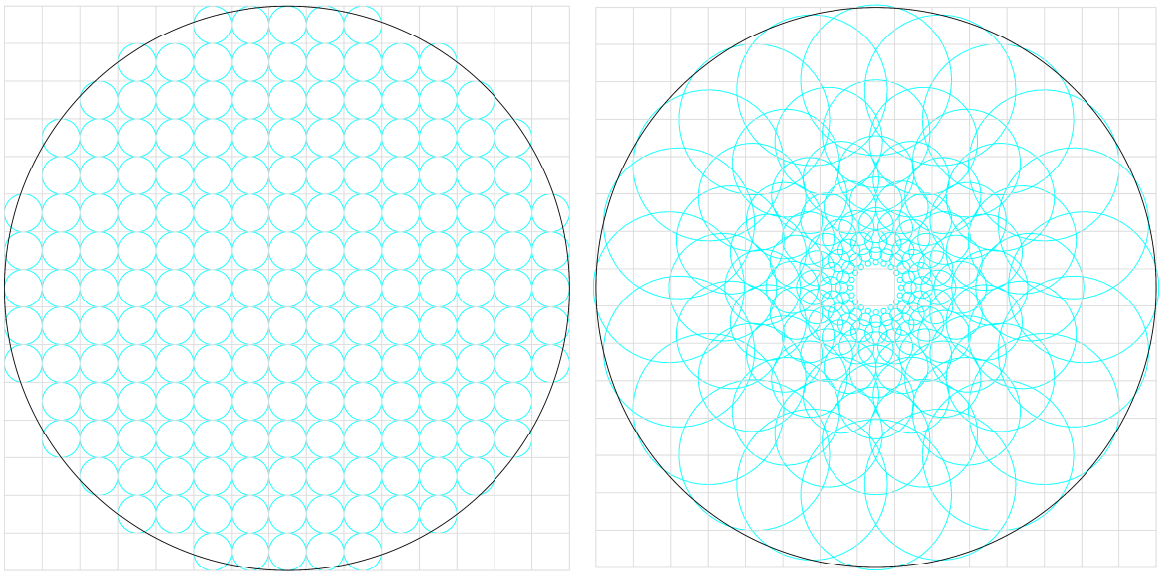
Let us consider that the cameras image planes $\mathcal{I}^l, \mathcal{I}^r \subset \mathbb{R}^2$ comprise a finite set of RFs denoted by $\mathcal{S}^l, \mathcal{S}^r \subset \mathbb{R}^2$. We assume that each RF has a non-uniform stimuli response, modeled by a two dimensional Gaussian profile [157], with the support regions depicted in Figure 3.6. The mean $\boldsymbol{\mu} = (\mu_x, \mu_y)$ defines the coordinates of the center of the RF in the retinal plane, where response is maximal, and the standard deviation σ represents its support radius.

Thus, observing a correspondence at a given RF pair $\mathbf{s}^i \in \mathcal{S}^l \times \mathcal{S}^r$ follows a conditional Gaussian distribution:

$$\mathbf{s}^i \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{s}^i}, \boldsymbol{\Sigma}_{\mathbf{s}^i}\right) \tag{3.6}$$

where

$$\boldsymbol{\mu}_{\mathbf{s}^i} = \begin{bmatrix} \mu_x^{l,i} \\ \mu_y^{l,i} \\ \mu_x^{r,i} \\ \mu_y^{r,i} \end{bmatrix} \quad , \quad \boldsymbol{\Sigma}_{\mathbf{s}^i} = \operatorname{diag}\left(\sigma^{l,i^2}, \sigma^{l,i^2}, \sigma^{r,i^2}, \sigma^{r,i^2}\right) \tag{3.7}$$



(a) Cartesian. (b) Log-polar.

Figure 3.6: Gaussian receptive fields with support plotted for 3 standard deviations

Stereoscopic Reconstruction

Given a point pair correspondence $\mathbf{s}^i$ found in the retinal domain, we determine the corresponding 3D position in the cyclopean reference frame via stereo analysis. Since point pair correspondences are inherently corrupted with precision errors, their projection lines may no longer satisfy the epipolar constraint and therefore not intersect in 3D space. Hence, one should rely on a triangulation method, denoted by τ, in order to compute a 3D Cartesian

point estimate $\hat{\mathbf{z}}$ from a point correspondence in image coordinates $\mathbf{s}^i$:

$$\tau : \mathcal{I}^l \times \mathcal{I}^r \longrightarrow \mathbb{R}^3 \tag{3.8}$$

Due to its simplicity and relatively low computational complexity, we use the mid-point method (for details please refer to [213]).

Uncertainty propagation via the Unscented Transform

Since the transformation (3.8) involved in 3D reconstruction is non-linear, we employ the Unscented transform [107] to compute the propagated mean and covariance up to the third order (by Taylor's expansion). This is achieved by approximating a multivariate Gaussian distributed variable with a set of meaningful and deterministically chosen set of samples (usually named sigma points). For each receptive field pair $\mathbf{s}^i \in \mathcal{S}^l \times \mathcal{S}^r$ we associate a set of sigma points

$$\mathcal{U}^i = \{\mathcal{X}^{(i,j)} \in \mathcal{I} \times \mathcal{I}' : j = 0, ..., 2N_s\} \tag{3.9}$$

where N_s is the number of sigma points. The sigma points are pre-computed according to the following expressions

$$\mathcal{X}^{(i,0)} = \boldsymbol{\mu}_{\mathbf{s}^i} \tag{3.10}$$

$$\mathcal{X}^{(i,j)} = \boldsymbol{\mu}_{\mathbf{s}^i} + \left(\sqrt{(N_s+\lambda)\boldsymbol{\Sigma}_{\mathbf{s}^i}}\right)_j \text{ for } j = 1, ..., N_s \tag{3.11}$$

$$\mathcal{X}^{(i,j)} = \boldsymbol{\mu}_{\mathbf{s}^i} - \left(\sqrt{(N_s+\lambda)\boldsymbol{\Sigma}_{\mathbf{s}^i}}\right)_j \text{ for } j = N_s+1, ..., 2N_s \tag{3.12}$$

where $(\cdot)_j$ denotes the j-th row of a matrix. Furthermore, we consider a set of weights

$$\mathcal{W} = \{w_c^{(j)}, w_m^{(j)} \in \mathbb{R} : j = 0, ..., 2N_s\} \tag{3.13}$$

which are computed as follows

$$w_m^{(0)} = \frac{\lambda}{L+\lambda} \tag{3.14}$$

$$w_c^{(0)} = \frac{\lambda}{L+\lambda} + \left(1 - \alpha^2 + \beta\right) \tag{3.15}$$

$$w_m^{(j)} = w_c^{(j)} = \frac{1}{2\left(L+\lambda\right)} \quad \text{for } j = 1, ..., 2N_s \tag{3.16}$$

where $\lambda = \alpha^2(L+K) - L$ is a scaling factor, α controls the spread of the sigma points around the mean, K is a secondary scaling parameter, and β is used to incorporate prior knowledge about the distribution of $\mathbf{s}$ (for Gaussian distributions $\beta = 2$ is optimal). Then, for a given point correspondence in retinal domain, we first apply the non-linear transformation τ to the sigma points associated with the corresponding RF pair, $\mathcal{Z}^{(i,j)} = \tau(\mathcal{X}^{(i,j)})$, and then re-estimate the mean and covariance in the 3D domain, according to

$$\hat{\boldsymbol{\mu}}_{\mathbf{z}^i} = \sum_{j=0}^{2N_s} w_m^{(j)} \mathcal{Z}^{(i,j)} \tag{3.17}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{z}^i} = \sum_{j=0}^{2N_s} w_c^{(j)} \left(\mathcal{Z}^{(i,j)} - \hat{\boldsymbol{\mu}}_{\mathbf{z}^i}\right)\left(\mathcal{Z}^{(i,j)} - \hat{\boldsymbol{\mu}}_{\mathbf{z}^i}\right)^T \tag{3.18}$$

The sigma points in retinal domain are computed offline and stored in a linear array in order to speed up on-line uncertainty propagation.

3.3.3 Randomized Sensory Ego-Sphere

In the proposed framework the SES plays an intermediate role between the stereo sensor and the decision planning process. Probabilistic data arriving from the sensory stream is continuously fused and integrated over time in the SES, by means of recursive Bayesian filtering. At the same time the available information is used to predict and redirect gaze to the best expected location, in the light of new observations.

Definition

The proposed SES is composed of a set of cells $\mathcal{P}$ lying on a unit sphere, oriented according to the world reference frame, and a map that assigns to each cell the 3D coordinates of the point observed by the robot at that orientation

$$\mathcal{M} : \mathcal{P} \longrightarrow \mathbb{R}^3 \tag{3.19}$$

The proposed cell grid structure is analogous to a Voronoi diagram defined on a spherical 2-manifold $\mathbb{S}^2$ in 3D space, as depicted in Figure 3.7. In practice the proposed SES comprises a set of $3D$ Cartesian sample points with unit norm and centered in the observer egocentric reference frame $\mathcal{E}$, aligned with the world reference frame

$$\mathcal{P} = \{\mathbf{p}^i \in \mathbb{R}^3, i, ..., N_x : \|\mathbf{p}^i\| = 1\} \tag{3.20}$$

which are i.i.d. and randomly generated from a three dimensional GMM distribution

$$\mathbf{p}^i = \frac{\mathbf{v}^i}{\|\mathbf{v}^i\|} \text{ where } \mathbf{v}^i \sim p(\boldsymbol{\theta}) = \sum_{m=1}^{M} \phi^m \mathcal{N}\left(\boldsymbol{\mu}_p^m, \boldsymbol{\Sigma}_p^m\right) \tag{3.21}$$

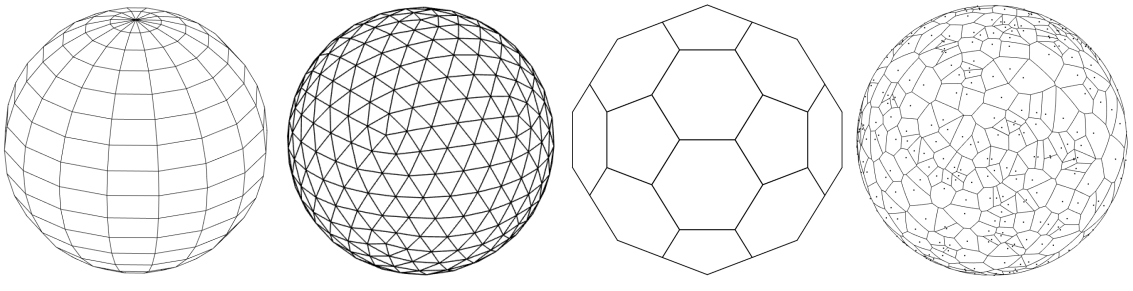
where M is the number of mixture components and where each $\mathbf{p}^i \in \mathcal{P}$ represents an orientation, allowing for efficient data-alignment with observed 3D points, using inner products (equation 3.23). Each SES cell, represented by $\mathbf{p}^i \in \mathcal{P}$, stores one environment sample point estimate $\mathbf{x}^i \in X$.

The statistics of the GMM distribution are chosen according to the observer goals. On one hand, in order to produce uniform and unbiased memory structures, the surface should be sampled from a rotationally symmetric distribution, i.e., from a single Gaussian with zero mean and variance equal in all dimensions [147] (Figure 3.7d). On the other hand, non-uniform, task-dependent memory biasing can be achieved by manipulating the GMM parameters, as can be seen in Figure 3.7. The proposed randomized representation offers a convenient mechanism for encoding task and world prior knowledge. Memory biasing should lead to more efficient, flexible and adaptable memory allocation and to more effective behaviours during task execution.

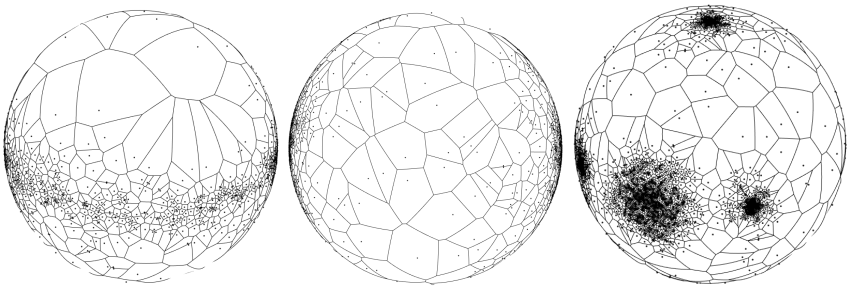
Hypothetical topologies that may be suitable for different tasks are depicted in Figure 3.7: If for instance the task is to look for people, one should privilege areas at the equator rather than the poles. In this case, varying the Gaussian mean is not sufficient. One could sample from a single-component zero mean GMM with larger variance in the horizontal directions (Figure 3.7e). Each SES cell (represented by the point $\mathbf{p}^i \in \mathcal{P}$) stores one environment sample point estimate $\mathbf{x}^i \in X$. While crossing a street, the observer should prioritize attentional resources to antipodal, lateral regions (Figure 3.7f). This can be achieved by sampling from a single-component Gaussian with a larger variance in the lateral component, or from a two-component GMM with opposite lateral means. More complex tasks can benefit from irregular topologies with multiple foci, obtained from GMMs with many components (Figure 3.7g).

Data Alignment

For each observed world point estimate provided by our stereo observation model at time t, $\mathbf{z}_t^o$, we need to find the associated memory cell in order to perform probabilistic data fusion. The association process goes as follows. First, the observed random variable is transformed from the cyclopean to the egocentric reference frame, according



(a) 2D array. (b) Geodesic dome. (c) Truncated icosahedron. (d) Unbiased.



(e) Equator-biased. ($M = 1$). (f) Antipodal. ($M \geq 1$). (g) Non-trivial ($M > 1$).

Figure 3.7: Different Sensory Ego-Spheres, resulting from different tessellations: top row illustrates highly regular, deterministic structures. The bottom row depicts our novel randomized structure for different task-dependent biases.

to the linear transformation $Z' : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ of the form

$$Z' = {}^{\mathcal{E}}\mathbf{R}_{\mathcal{C}} Z + {}^{\mathcal{E}}\mathbf{t}_{\mathcal{C}} \tag{3.22}$$

where ${}^{\mathcal{E}}\mathbf{R}_{\mathcal{C}} \in \mathbb{R}^{3\times3}$ is an orthogonal rotation matrix and ${}^{\mathcal{E}}\mathbf{t}_{\mathcal{C}} \in \mathbb{R}^{3\times1}$ a translation vector, obtained by combining proprioceptive joint angle measurements with known forward kinematics.

Second, for each observation $\mathbf{z'}_t^o$ we find the associated memory cell c^o, which is the one that minimizes the Euclidean distance, according to the mapping function $\mathcal{M}$, here defined as follows

$$c^o = \mathcal{M}\left(\hat{\boldsymbol{\mu}}_{\mathbf{z'}_t^o}\right) = \underset{j}{\operatorname{argmin}} < \mathbf{p}^j, \frac{\hat{\boldsymbol{\mu}}_{\mathbf{z'}_t^o}}{\|\hat{\boldsymbol{\mu}}_{\mathbf{z'}_t^o}\|} > \tag{3.23}$$

After finding the associated cell we update its respective estimate according to equation (3.33). Moreover, we assume that the transformed observations are conditionally independent, given X_t, and thus

$$p(Z'_t|X_t) = \prod_{o=1}^{N_v} p(\mathbf{z'}_t^o|\mathbf{x}_t^{c^o}) \tag{3.24}$$

Finally, the resulting probabilistic observation model $p(\mathbf{z'}_t^o|\mathbf{x}_t^{c^o})$ follows a Gaussian distribution

$$\mathbf{z'}_t^o|\mathbf{x}_t^{c^o} \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_{\mathbf{z'}_t^o}, \hat{\boldsymbol{\Sigma}}_{\mathbf{z'}_t^o}\right) \tag{3.25}$$

with statistics computed as follows

$$\hat{\boldsymbol{\mu}}_{\mathbf{z'}_t^o} = {}^{\mathcal{E}}\mathbf{R}_{\mathcal{C}} \hat{\boldsymbol{\mu}}_{\mathbf{z}_t^o} + {}^{\mathcal{E}}\mathbf{t}_{\mathcal{C}} \tag{3.26}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{z'}_t^o} = {}^{\mathcal{E}}\mathbf{R}_{\mathcal{C}} \hat{\boldsymbol{\Sigma}}_{\mathbf{z}_t^o} \mathbf{R}_{\mathcal{C}}^T \tag{3.27}$$

Probabilistic Sensor Fusion

In the sensor fusion perspective, the goal of the optimal Bayesian estimator is to determine the posterior probability distribution over X, given the accumulated visual sensory observations and the robot proprioceptive state measurements up to time $t \in \mathbb{N}$. Sequential Bayesian filtering allows us to accumulate sensor inputs and update the likelihood of X, at each time instant.

The posterior probability distribution at time t, of the set of internal environment sample points X_t given the current and past visual and proprioceptive observations, is given by

$$p(X_t|Z_{1:t}, U_{1:t}) = p(X_t|Z_t, Z_{1:t-1}, U_{1:t}) \tag{3.28}$$

Furthermore, since we assume that the proprioceptive measurements are deterministic, then

$$p(Z'_t|Z_t, U_t) = p({}^{\mathcal{E}}\mathbf{R}_{\mathcal{C}} Z_t + {}^{\mathcal{E}}\mathbf{t}_{\mathcal{C}}) = p(Z'_t) \tag{3.29}$$

and equation (3.28) becomes

$$p(X_t|Z'_{1:t}) = p(X_t|Z'_t, Z'_{1:t-1}) \tag{3.30}$$

Since the world is static, at each iteration, the solution to the filter involves only one update step: in the *measurement update* step observations are used to update the current belief by applying the Bayes rule to the right hand side of equation (3.30) and using the observation model (3.5) we get

$$p(X_t|Z'_{1:t}) = \eta p(Z_t|X, Z'_{1:t-1}) p(X|Z'_{1:t-1}) \tag{3.31}$$

where η is a normalizing constant. Since the current observations Z'_t are conditionally independent of the past observations $Z'_{t:t-1}$ given the current environment projection in the egocentric frame, X_t, the previous equation becomes

$$p(X_t|Z'_{1:t}) = \eta p(Z'_t|X_t) p(X_t|Z'_{1:t-1}) \tag{3.32}$$

The *a posteriori* is independently determined for each cell, according to

$$p(\mathbf{x}_t^{c^o}|\mathbf{z'}^o_{1:t}) = \eta p(\mathbf{z'}^o_t|\mathbf{x}_t^{c^o}) p(\mathbf{x}_t^{c^o}|\mathbf{z'}^o_{1:t-1}) \tag{3.33}$$

and follows a Gaussian distribution, with statistics given by

$$\hat{\mathbf{\Sigma}}_{\mathbf{x}_t^{c^o}} = \left(\hat{\mathbf{\Sigma}}^{-1}_{\mathbf{x}_{t-1}^{c^o}} + \hat{\mathbf{\Sigma}}^{-1}_{\mathbf{z'}_t^o} \right)^{-1} \tag{3.34}$$

$$\hat{\boldsymbol{\mu}}_{\mathbf{x}_t^{c^o}} = \hat{\mathbf{\Sigma}}_{\mathbf{x}_t^{c^o}} \left(\hat{\mathbf{\Sigma}}^{-1}_{\mathbf{x}_{t-1}^{c^o}} \hat{\boldsymbol{\mu}}_{\mathbf{x}_{t-1}^{c^o}} + \hat{\mathbf{\Sigma}}^{-1}_{\mathbf{z'}_t^o} \hat{\boldsymbol{\mu}}_{\mathbf{z'}_t^o} \right) \tag{3.35}$$

Each point estimate is initialized with large mean and covariance to reflect the high uncertainty due to non-existent world prior knowledge.

3.3.4 Active Vision: Sequential Stochastic Decision Making

In the proposed framework, the decision making is responsible for sensory-motor coordination. Based on probabilistic information stored in memory, the decision process selects where to look next and the associated desired motor commands. Like other approaches that use uncertainty and task-related rewards to guide decision making we frame our approach within the reinforcement learning domain [196]. As such, the agent selects the

action that maximizes expected task related cumulative rewards.

The underlying framework for decision making under uncertainty, which assumes non-deterministic noisy state observations, is known as POMDPs. In our particular problem formulation we have continuous states and observations, as well as large discrete action spaces (i.e. each SES cell represents an action), which renders intractable the computation of optimal policies. Even approximate methods for solving POMDPs would take considerable time (e.g. hours or days). Moreover, if the environment changes the observer needs to recompute the full policy. Hence, state-of-the-art methods for solving POMDPs are unsuitable for large problems, which require real-time on-line decision making. Therefore, rather than framing our problem as a POMDP and relying on the computation of Bayesian optimal policies which are typically intractable for large state spaces, we rely on simpler and less costly tools from Bayesian Optimization, for reinforcement learning. More concretely, from Multi-Armed Bandit (MAB) [174].

Saccadic Planning as a Multi-armed bandit Problem

In MAB problems, at each time instant the agent selects an action and collects a reward. The rewards are drawn from a posterior probability distribution whose statistics are continuously updated over time. Typically, the goal of the agent is to maximize the sum of collected rewards or, equivalently, minimize cumulative regret. In this work the selected task was to find the closest object to the observer as fast (i.e. with minimum fixations) and precisely (i.e. with minimum uncertainty) as possible. Within the MAB framework, this is commonly referred to as the best-arm identification problem [9]. In our particular setting, each world sample point represented in memory is a bandit whose statistics are not known in advance. The agent chooses actions, i.e. a fixation point, from the set of alternatives $a \in \{1, ..., N_x\}$ and collects payoffs from a reward distribution $r(\mathbf{x}^a)$. Considering the task at hand, we define the reward obtained when choosing a given action a as a function of the distance to the ego-frame

$$r(\mathbf{x}^a) = -\|\mathbf{x}^a\|_2 \tag{3.36}$$

Since $\mathbf{x}^a$ follows a Gaussian distribution, then we consider a first order approximation for the reward distribution such that

$$r(\mathbf{x}^a) \sim \mathcal{N}\left(\mu_r(\mathbf{x}^a), \sigma_r(\mathbf{x}^a)\right) \tag{3.37}$$

where $\mu_r(\mathbf{x}^a)$ and $\sigma_r(\mathbf{x}^a)$ are computed as follows

$$\mu_r(\mathbf{x}^a) = E\left[-\|\mathbf{x}^a\|_2\right] = -\|\hat{\boldsymbol{\mu}}_{\mathbf{x}^a}\|_2 \tag{3.38}$$

$$\sigma_r(\mathbf{x}^a) = \mathrm{Var}\left[-\|\mathbf{x}^a\|_2\right] \approx \mathbf{J}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{x}^a} \mathbf{J} \tag{3.39}$$

where $E[\cdot]$ and $\mathrm{Var}[\cdot]$ denote the expectation and variance operators, respectively, and $\mathbf{J}$ is a Jacobian matrix, defined as follows

$$\mathbf{J} = \left.\frac{\partial r(\mathbf{x})}{\partial \boldsymbol{x}}\right|_{\hat{\mu}_{\mathbf{x}^a}} = \left[\frac{x_{\hat{\mu}_{\mathbf{x}^a}}}{\|\hat{\mu}_{\mathbf{x}^a}\|_2} \frac{y_{\hat{\mu}_{\mathbf{x}^a}}}{\|\hat{\mu}_{\mathbf{x}^a}\|_2} \frac{z_{\hat{\mu}_{\mathbf{x}^a}}}{\|\hat{\mu}_{\mathbf{x}^a}\|_2}\right]^T \tag{3.40}$$

Acquisition Functions

In the Bayesian optimization framework, acquisition functions are responsible for defining the strategy when searching for the optimum. The literature on acquisition functions used to guide stochastic optimization is vast and includes many different heuristics that deal with the exploration-exploitation dilemma. On one hand, Probability of Improvement (PI) [119] methods select the action that maximizes the probability of improving the current

instantaneous reward. On the other hand, Expected Improvement (EI) [142] seeks for the action that maximizes the expected improvement magnitude. More recently, the idea of using UCBs [120] to deal with exploration-exploitation trade-offs in machine learning problems has proven successful in robotics applications [126], exhibiting increased preference for exploration when compared to the former approaches. Since the best performing acquisition function is highly dependent on the objective at hand, the authors in [88] propose combining single acquisition functions in mixed portfolio strategies.

In this work we compared three different action selection strategies:

1. a simple yet powerful UCB algorithm named "Sequential Design for Optimization" [40] that is easy to implement and elegantly handles the trade-off between exploration (minimizing uncertainty) and exploitation (maximizing rewards) that emerges in decision making under uncertainty [3]. At each time instant, the observer selects the alternative with maximal upper confidence bound on the expected reward, given the past observations, according to the following expression

$$a_t = \underset{a \in \{1,\ldots,N_x\}}{\operatorname{argmax}} \; \mu_r(\mathbf{x}_t^a) + \alpha \sigma_r(\mathbf{x}_t^a) \tag{3.41}$$

   where α is a user selected parameter that controls the width of the confidence bound and thus the exploration behaviour during task execution.

2. The probability of improvement, which at each time instant, selects the action with highest probability of leading to an improvement upon the current best ($\mathbf{x}_t^*$), as follows

$$a_t = \underset{a \in \{1,\ldots,N_x\}}{\operatorname{argmax}} \; \mathbb{P}(r(\mathbf{x}_t^a) > r(\mathbf{x}_t^*)) \tag{3.42}$$

3. The expected improvement, which tries to maximize the expected magnitude of the improvement upon the so far best, according to

$$a_t = \underset{a \in \{1,\ldots,N_x\}}{\operatorname{argmax}} \; \mathbb{E}(r(\mathbf{x}_t^a) - r(\mathbf{x}_t^*)) \tag{3.43}$$

$$\tag{3.44}$$

Finally, the motor-action U_t corresponding to $\mathbf{x}_t^{a_t}$ is computed from known forward kinematics.

3.4 Experiments and Results

In order to demonstrate the applicability of the proposed framework and compare the performance of different visual sensor topologies, we performed a set of experiments in simulation. In all of the experiments we constrained the number of RFs - and hence the computational resources - to be always fixed and equal in the Cartesian and log-polar cases (please refer to [157] for mathematical details on the log-polar distribution). We considered $N_{rf} = 200 \times 200$ images in both cases.

The remainder of this section is organized as follows. We begin by characterizing and assessing the ability of the different sensors to map the environment with low uncertainty. Then, we proceed to evaluating the performance of the complete active task-oriented stereo sensing framework, in a realistic simulated environment.

3.4.1 Sensor Characterization

To characterize the proposed sensor model, we assessed the average uncertainty in 3D reconstruction as a function of depth, vergence angle and sensor type in the following manner: First, we generated a set of fronto-parallel planar surfaces, with varying distance $d \in [0, 1]$ from the binocular system. Depth is constant for all points lying within the same planar-surface. Then, for each planar surface we varied the vergence angle, in the interval $\theta^v \in \left[0, \frac{\pi}{2}\right]$ and computed the corresponding 3D reconstructions with associated uncertainties. Note that in this experiment we are not characterizing a full environment but just single snapshots taken by the observer. Furthermore, we assumed that an object can be approximated by a planar surface occupying the observing agent field of view. This allows for comparing both sensors, under the same conditions.

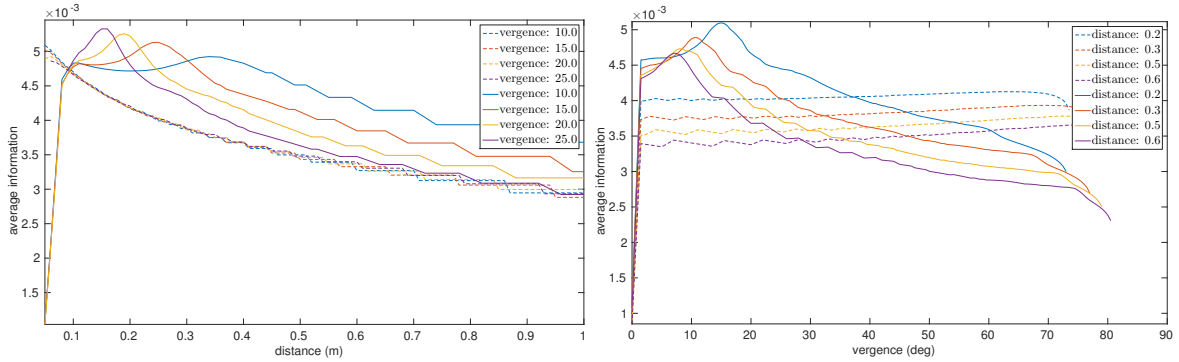
Let us consider the log-determinant of the inverse covariance matrix (also known as precision matrix) to quantify pointwise information:

$$I\left(\boldsymbol{\Sigma}\right) = -\log(|\boldsymbol{\Sigma}|) \tag{3.45}$$

Here we rely on the average information gathered with a single depth image to assess the quality of the sensors, which is defined as folllows

$$TI = \frac{1}{N_{rf}} \sum_{i=1}^{N_{rf}} I\left(\hat{\boldsymbol{\Sigma}}^i\right) \tag{3.46}$$

As depicted in Figure 3.8, the foveal outperforms the Cartesian sensor, in terms of gathered information which is maximal if the fixation point coincides with the planar surface. Furthermore, the Foveal sensor information reliability decays monotonically with increasing depth, and is more dependent on the vergence angle than on the Cartesian one.. These results are directly in line and support previous findings [216] that suggest that foveal distributions facilitate stereo vision in convergent systems. In foveated systems gaze acts like a focus of attention, which when directed to the point of interest, improves dramatically the depth resolution around the fixation point. Instead, optical vergence movements in Cartesian systems provide no gains in 3D resolution, resulting only in unnecessary energetic costs.



(a) (b)

Figure 3.8: Numerical characterization of the sensor model for the Cartesian (dashed lines) and the Log-polar (solid lines) sensors, as a function of distance and vergence. (a) Varying distance for diferent vergence angle curves. (b) Varying vergence for different planar distance curves.

3.4.2 Active Vision

With the view of investigating how our methodology performs in simultaneous target searching and mapping, we performed a set of experiments in the Gazebo simulator with the Vizzy robot [144] head (see Figure 3.9). For the

(a) (b)

Figure 3.9: The simulation scenario created for evaluating the proposed active vision framework. The task to perform was to find the nearest object from the robot ego frame. The evaluation scenario contained a non-trivial global optimum which could only be attended if enough exploration was promoted. (a) The simulation scenario created for evaluating the proposed active vision framework. (b) The global optimum was placed at a non-trivial location which could only be attended with either sufficient exploration or a wide field of view.

sake of the experiments simplicity, the robot was fixed to the ground floor and hence the motion was restricted to oculocephalic movements. However, note that our methodology is also applicable to scenarios in which the robot platform can move. This would imply updating the 3D point estimates stored in memory taking into account the uncertainty in robot base movements (odometry), and implementing a z-buffer technique to determine which point to store in each cell, due to possible occlusions occurring after translations.

We created a static scenario with multiple objects (coke cans) displaced at arbitrary depths, in which a set of objects (coke cans) were strategically displaced over a highly textured background, in order to facilitate stereo reconstruction, which is highly dependent on the environment texture richness. The Gazebo simulator generates pinhole camera images, with uniform resolution. Hence, for the log-polar sensor, we generated foveated images from uniform resolution images by first applying the log-polar transformation and then converting back to Cartesian domain via the inverse transformation. This operation has the effect of blurring the image in the periphery while maintaining high resolution in the center. Finally, disparity maps were computed using a state-of-the-art dense stereo matching algorithm named Semi-Global Block Matching (SGBM) [87].

As previously pointed out, the task at hand was to find the nearest world point to the observer. Points on the ground floor are easily excluded by thresholding the z_w coordinate. In all experiments we fixed the number of memory sample points to $N_x = 20000$. In each experiment we let the observer perform $T = 50$ saccadic movements, with initial ($t = 1$) pan, tilt and vergence angles equal to zero. Each experiment was repeated 20 times in order to average out variability in different real-time simulations. Non-repeatability was influenced by multiple factors including separate threads for Gazebo's physics and sensor generation, as well as stochastic delays involved in higher level inter-process communication. Furthermore, in order to deal with motion blur and visuo-proprioceptive delays that arise during saccadic eye movements, we used the visual suppression mechanism proposed in [11], which temporarily blinds the observer during saccades.

Evaluation Metrics

In order to quantitatively assess the performance of our methodologies we considered the following evaluation metrics:

- the gap reduction metric [93] which is a quality measure that evaluates how effectively the algorithm is at

finding the global maximum:

$$g_t = \frac{\mu_r(\mathbf{x}^+) - \mu_r(\hat{\mathbf{x}}_1^{a_1})}{\mu_r(\mathbf{x}^*) - \mu_r(\hat{\mathbf{x}}_1^{a_1})} \tag{3.47}$$

where $\mu_r(\mathbf{x}^*)$ is the true global maximum

$$\mu_r(\mathbf{x}^*) = \max_i \mu_r(\mathbf{x}^i) \tag{3.48}$$

and $\mu_r(\mathbf{x}^+)$ is the best obtained reward up to time t

$$\mu_r(\mathbf{x}^+) = \max_t \mu_r(\hat{\mathbf{x}}_t^{a_t}) \tag{3.49}$$

The gap is defined between 0, meaning no improvement over the initial fixation, and 1 for the optimal improvement. In order to measure the speed for task completion and thus performance efficiency we also assess the average gap reduction per saccade which implicitly represents the average progress towards the optimum per saccade:

$$G/S = \frac{1}{T}\sum_{t=1}^{T} g_t \tag{3.50}$$

- the cumulative regret which is a standard metric, here suitable to evaluate the convergence behaviour during the search for the optimum:

$$R_t = \mu_r(\mathbf{x}^*) - \frac{1}{t}\sum_{k=1}^{t} \mu_r(\hat{\mathbf{x}}_k^{a_k}) \tag{3.51}$$

Notice that here we are not interested in minimizing the total regret, i.e. the incurred losses during exploration, but instead on finding the global optimum.When normalized by the number of saccades it represents the temporal cumulative regret gain per saccade:

$$R/S = \frac{1}{T}\sum_{t=1}^{T} R_t \tag{3.52}$$

- the average global gathered information which is a quality performance measure of the global knowledge gathered about the world up to time t (exploratory behaviour):

$$GI_t = \frac{1}{N_x}\sum_{i=1}^{N_x} I\left(\hat{\mathbf{\Sigma}}_t^i\right) \tag{3.53}$$

When normalized by the number of saccades it represents the temporal average global information gain per saccade:

$$GI/S = \frac{1}{T}\sum_{t=1}^{T} GI_t \tag{3.54}$$

- the nearest object gathered information, which is a target reconstruction quality measure that benefits high

precision (i.e. low uncertainty) in target reconstruction:

$$LI_t = I\left(\hat{\mathbf{\Sigma}}_t^i\right) \quad \forall_{i:\|\hat{\boldsymbol{\mu}}_{\mathbf{x}_t^*} - \hat{\boldsymbol{\mu}}_{\mathbf{x}_t^i}\| < R_{NN}} \tag{3.55}$$

where $\hat{\mathbf{\Sigma}}_t^*$ is the true known global maximum estimated covariance at time t and R_{NN} is a user-selected nearest-neighbor radius. We considered $R_{NN} = 0.1m$ in all the experiments described below.

When normalized by the number of saccades it represents the temporal average local information gain per saccade:

$$LI/S = \frac{1}{T}\sum_{t=1}^{T} LI_t \tag{3.56}$$

Foveal vs Cartesian

Our first aim was to compare the behaviour of the foveal against the Cartesian sensor during task execution, for different upper confidence bound parameter values $\alpha \in \{0, 0.01, 1, 100, \infty\}$ and different sensing field of views fov $\in \{90ž, 135ž\}$. The sensor field of views were selected such that in one of the cases (fov $= 90ž$) the global optimum was not in the field of view of the observer at $t = 1$. The SES cells were generated from a unbiased, zero mean Gaussian distribution at initialization (see Table 3.1).

A global analysis of the results depicted in Figure 3.11 shows that the foveal sensor outperforms the Cartesian both in terms of the quality of the gathered information, as well as the task execution speed and effectiveness, as demonstrated by the gap reduction plot. We hypothesize that the best performance of the foveal sensor is due to the fact that the uncertainty in the periphery implicitly promotes more peripheral (lateral) exploration whereas the Cartesian promotes longitudinal (depth) search. This statement is clearly supported by the cumulative regret plots which exhibit lower losses for the Cartesian sensor, and thus a greedier behaviour. Moreover, for the foveal sensor case, a larger FOV allows the agent to attend the target more quickly at the cost of reduced information gain. A wider FOV, despite having less peripheral resolution, is advantageous in the speed of execution during visual search tasks.

In Figure 3.11a we assess the performance of our method for the different acquisition functions referred in section 3.3.4. On one hand, in the UCB case, a larger confidence bound parameter α increases exploration and, on average, improves performance in the particular task of finding the nearest object. However, too much exploration incurs in large cumulative regrets, and thus in high energy costs due to large oculocephalic movements incurred when attending objects further from the observer. Nevertheless, purely exploratory behaviours ($\alpha = \infty$) lead to better results in the average reconstruction quality as shown by the information metrics, since on average more memory sample locations are fixated. On the other hand, the tested improvement-based policies (PI and EI) seek to improve on the current best and have the advantage of being parameter free. For our particular setting, and similarly to UCB with $\alpha = 0$, PI tends to be excessively greedy and get trapped in local minima. On the contrary, EI deals well with the exploration-exploitation trade-off, as demonstrated by the average gap reduction and cumulative regret per saccade metrics due to the fact that it implicitly accounts for the improvement magnitude of each saccadic action, which allows for choosing distant, with high variance, fixation points.

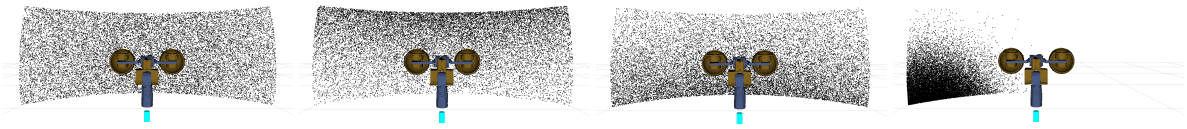
An in-depth analysis of the temporal evolution metrics (Figure 3.11b) for a fixed $\alpha = 100$, allows us to assess convergence times for a fairly exploratory behaviour. The temporal evolution of the gap reduction metric shows that, in all cases, no more than 20 saccades are necessary to perform the task of finding the nearest object for both sensor types. Howbeit, as indicated by the accumulated regret temporal evolution, convergence is only achieved after no less then 30 saccades. We further note that, after convergence, the cumulative regret is on average higher for the foveal case, as a consequence of having a more exploratory nature. Other than the Cartesian sensor with a FOV of $135ž$, all cases were successful on average in the task of finding the nearest object and did not get trapped in local minima. We believe that the poor performance of the Cartesian sensor, in this particular setting, is a result

Bias	$\boldsymbol{\mu}_p$			$\boldsymbol{\Sigma}_p$		
	x	y	z	xx	yy	zz
unbiased	0	0	0	0.5	0.5	0.5
top	0	0	1	0.5	0.5	0.5
down	0	0	-1	0.5	0.5	0.5
target	0.61	0.43	-0.67	0.05	0.05	0.05

Table 3.1: Memory biasing parameters.

of prioritizing points that are further from sensor, and not necessarily in the periphery, where the optimum lies. The gathered information exhibits an asymptotically convergence behaviour and has a faster transient time for the Cartesian sensor, again, supporting the idea that the Cartesian sensor is more greedy, myopic, and thus more prone to get trapped in local minima. We further note that the average cumulative regret is on average lower for the Cartesian case, again, as a consequence of having a more exploratory nature.

Memory Biases



(a) Uniform. (b) Top bias. (c) Bottom bias. (d) Target bias.

Figure 3.10: SES sample point distribution according to different topological memory biases and kinematic constraints.

Here our goal was to investigate the effect of different spatial memory topological biases imposed from *a priori* knowledge regarding the environment structure and the task at hand. At the present, experiments were performed with a foveal sensor with fov $= 135°$, and for the UCB with $\alpha = 100$. We intended to demonstrate that a careful displacement of the memory patches considering prior knowledge about the surrounding environment and the task at hand should incur in large performance gains. Therefore, we considered four different prior belief distributions with parameters defined in Table 3.1 and resulting SES topologies depicted in Figure 3.10:

- a "neutral" (unbiased) distribution reflecting the absence of a priori knowledge about the target location.
- a "bad" (top) prior belief distribution based on the wrong assumption that the object is above the observer.
- a "good" (down) prior belief distribution that assumes that the object is on the ground
- a "very good" (target) prior belief considering the true location of the target object.

In the Figure 3.12 we can observe that the "target" case had the best performance and the "top" the worst performance according to all metrics. In fact, as demonstrated by the gap reduction and the accumulated regret time evolution plots, the method was successful in finding the global optimum and converged with only 2 saccades. All the other cases were still able to find the optimum with less than 10 saccades and converge to the optimum within the first 20 saccades.

As expected, the gathered average local information metric indicates that increasing the memory sample density around the object of interest improves the target's gathered information. These experiments demonstrate that translating task-related priors in clever memory allocation to regions of higher reward yields faster task execution times and faster convergence rates. This results in an increase in the time spent on reducing the uncertainty on the target and therefore in improved reconstruction quality. On the one hand, promoting higher resolution in spatial memory to the most important surrounding regions according to the task, allows for more accurate target reconstruction. On the other hand, less fixations are needed to find the target, since less memory cells, and thus possible fixations, will reside outside of the target vicinity.

3.5 Conclusions

In this work we investigated the impact of uncertainty due to quantization phenomena in the retina and on how to take advantage of it to guide gaze shifts for two distinct retinal topologies: Cartesian and log-polar. With our approaches different gaze patterns emerge depending on the sensor topology and field of view and on exploration-exploitation confidence bounds parameters. The obtained results demonstrate that a wider field of view, despite less peripheral resolution is advantageous in visual search tasks execution speed. Furthermore, we showed that a task-biased SES allows for simultaneously coping with limited memory resources (i.e. limited number of memory cells) while improving performance, both in terms of target reconstruction quality and task execution speed.

We have proposed in **P.III** a novel visual search framework for robotic systems provided with binocular foveal vision. Our framework combines visual information and associated probabilistic 3D measures that are used to autonomously drive the agent's gaze direction during search tasks. Our contributions are the following:

1. First, we model the stereo reconstruction uncertainty that arises as a result of spatial quantization phenomena inherent in the retina. Our approach considers Gaussian RFs distributed in space following two different tessellations: (i) a classical uniform (Cartesian) arrangement and (ii) a log-polar one that mimics the human retina. The RFs in the latter present a space-variant spatial distribution and support radius.

2. The UT is used to propagate belief from the 2D retina domain to 3D via stereo reconstruction. When compared with previous approaches that also assume Gaussian quantization noise and that rely on first order linearizations to approximate the non-linear transformations involved in 3D reconstruction [115], our method is based on the more precise, third order approximations of the UT.

3. Probabilistic sensory measurements are integrated in a novel versatile randomized SES whose topology, unlike previously proposed structures in the literature, may be biased according to the autonomous agent short-term tasks and goals. The proposed SES, helps achieving the desired search goal, by allocating the limited memory resources more densely to important egocentrically encoded directions, according to the task while allowing for continuous sensory fusion via Bayesian estimation. The method proposed for generating a SES is based on projecting in the sphere randomly generated points according to a mixture of 3D Gaussians of arbitrary number of components, focal points (means) and dispersions (covariances). This generates an irregular grid but one can define more freely areas on the sphere with varying degrees of density and dispersion. The proposed randomized representation offers a convenient mechanism for encoding task and world prior knowledge. Memory biasing leads to more efficient, flexible and adaptable memory allocation and to more effective behaviours during task execution. Hypothetical topologies that may be suitable for different tasks are depicted in Figure 3.13: If for instance the task is to look for people, one should privilege areas at the equator rather than the poles. In this case, varying the Gaussian mean is not sufficient. One could sample from a single-component zero mean GMM with larger variance in the horizontal directions (Figure 3.13b). While crossing a street, the observer should prioritize attentional resources to antipodal, lateral regions (Figure 3.13c). This can be achieved by sampling from a single-component Gaussian with a larger variance in the lateral component, or from a two-component GMM with opposite lateral means. More complex tasks can benefit from irregular topologies with multiple foci, obtained from a GMM with many components (Figure 3.14b).

4. Finally, the decision-making process is framed within a MAB setting, that seeks to maximize expected task-related rewards. The proposed decision algorithm relies on statistical measures to decide where to look next by selecting the most promising regions to attend. We tested different stochastic optimization techniques to deal with the exploration-exploitation trade-off typical of decision making under uncertainty algorithms.

   With our approaches different gaze patterns emerge depending on the sensor topology and field-of-view and on exploration-exploitation confidence bound parameters. The obtained results demonstrate that a wider

field-of-view, despite less peripheral resolution is advantageous in task execution speed (see Figure 3.14a). Furthermore, we showed that a task-biased SES allows for simultaneously coping with limited memory resources (i.e. fixed number of memory cells) while improving performance, both in terms of target reconstruction quality and speed of execution (see Figure 3.14a).

Future work The main limitations of the proposed framework for 3D visual search tasks are due to the fact that they solely rely on depth information and associated uncertainty measures. One could benefit from RGB cues to incorporate recognition abilities and speed-up localization. The framework can be further enhanced with other ideas from the literature, namely:

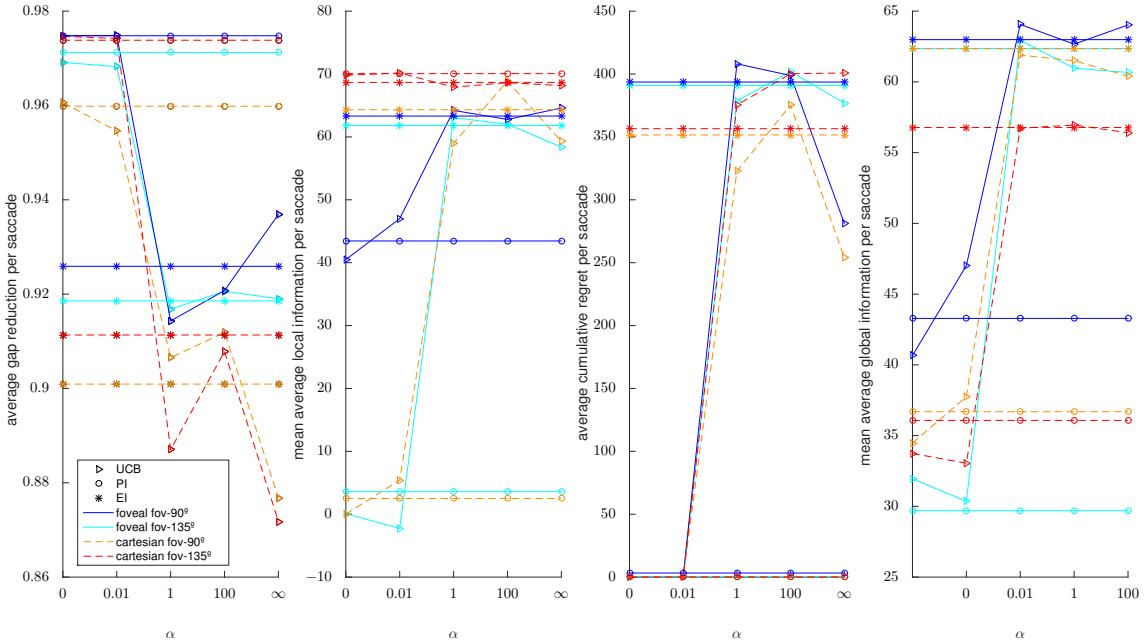
- *Multi-modal visual search for improved target 3D reconstruction accuracy:* applying the previously proposed foveal vision with DCNNs framework for RGB object dection, as a prior step to improve target localization, 3D local reconstruction, pose estimation accuracy, and to speed-up visual search tasks. Software integration and experiments would be prepared and designed to run on a real robotic platform available in the laboratory (see Figurere 1.1). Experiments could include assessing the performance on a visual-search, reconstruction and pose estimation task, in a robot grasping and manipulation application.

- *Faster memory access:* one could improve the proposed SES run-time performance by implementing a nearest neighbour data alignment scheme. A kd-tree could be built during initialization time and used for storing the ego-sphere cells. Then, for each observed 3D point, searching for the closest cell on the sphere would be performed in $\mathcal{O}(log(N_x))$, instead of $\mathcal{O}(N_x)$. Also, reconstruction quality and efficiency can be further improved by restricting stereo matching to biologically plausible volumes of interest [2].

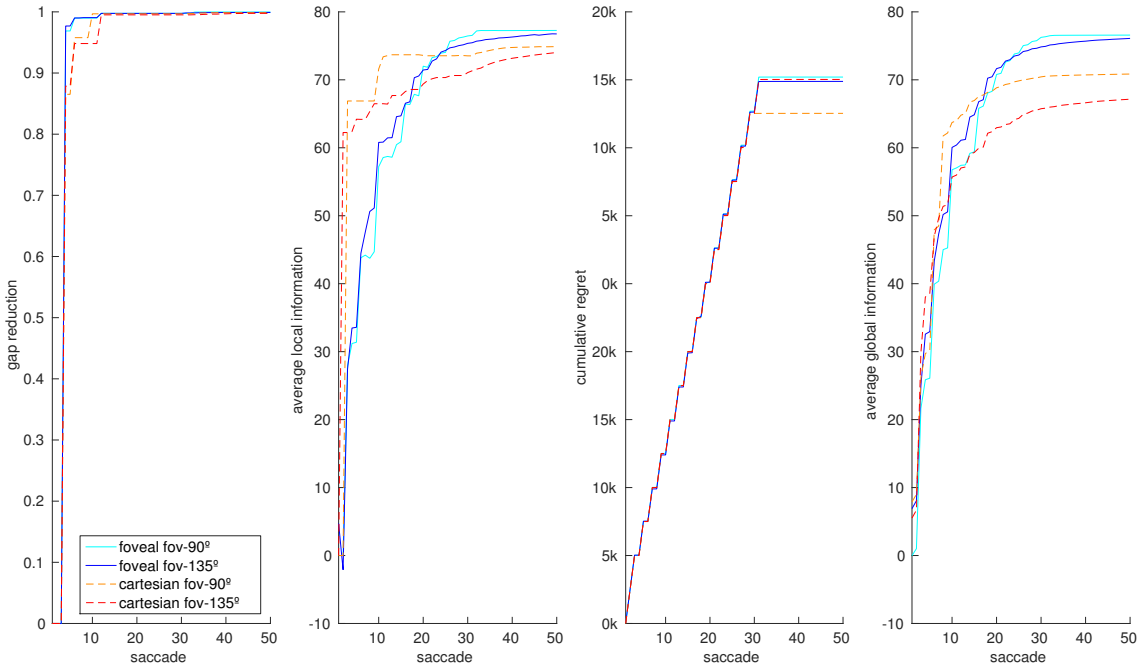
- *Dynamic spatial memory:* one limitation of the proposed SES stems from the fact that it is static, and defined before run-time. Adaptive re-sampling techniques (e.g. particle filters [8]) may be beneficial for improved target reconstruction accuracy, and dealing with dynamic tasks and environments (e.g. locating moving targets).

- *Egocentric planning and allocentric mapping:* another drawback of the current framework is the use of a SES for mapping the surrounding environment, which requires constant expensive updates as the observing agent moves, and the implementation of z-buffers to deal with occlusions and state's partial observability. Instead, in the future one can combine the benefits of the proposed egocentric flexible SES structure for NBV planning, with the more appropriate allocentric representations, e.g. the memory efficient voxel-based OctoMap structure [90], for environment mapping, path planning and navigation.

- *Learning task-dependent priors:* currently the most important directions for a specific task are encoded using GMM priors whose parameters are manually provided by humans. Instead one could directly learn task-specific priors, from data gathered from human demonstrations.

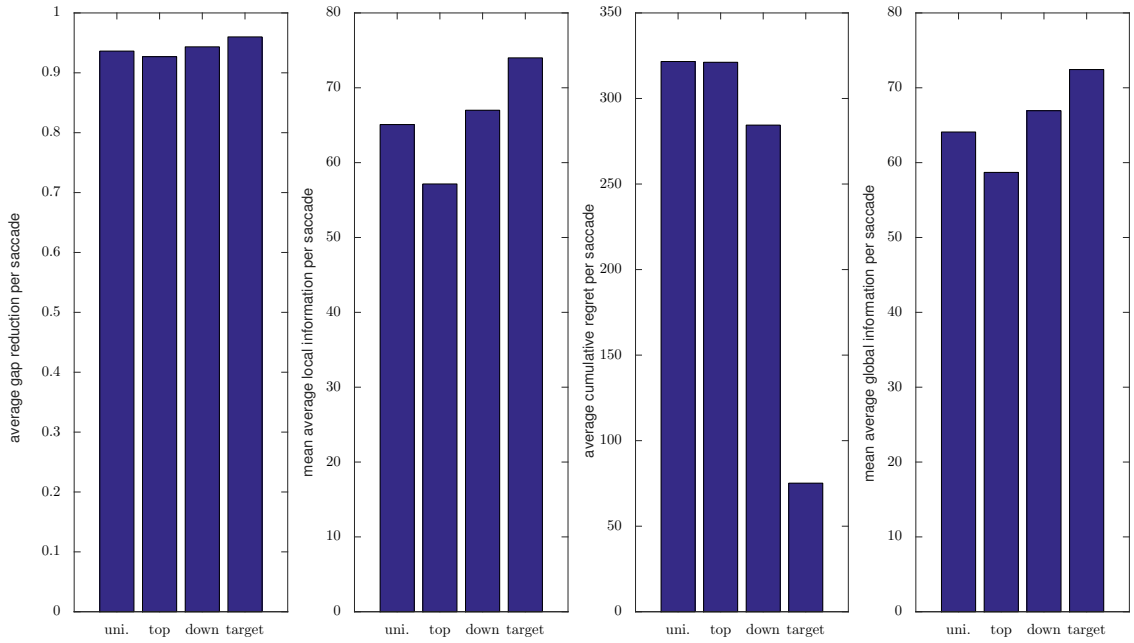


(a) Average per saccade performance plots.

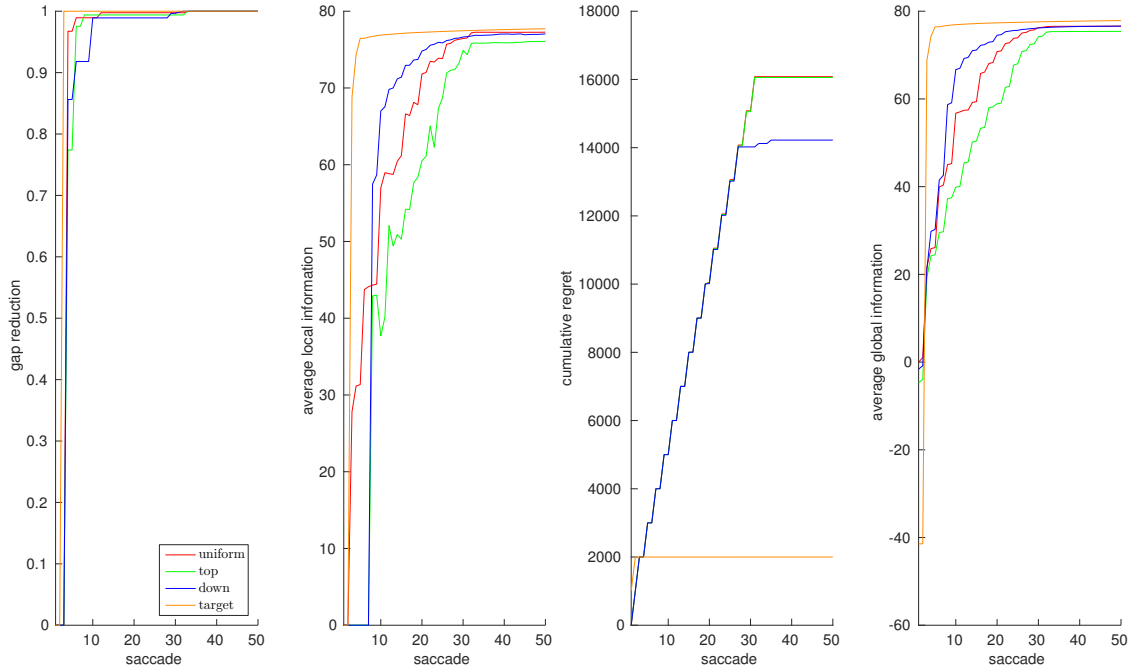


(b) Time evolution plots.

Figure 3.11: Performance results for the assessed sensor topologies, field of views and upper confidence bound parameter.

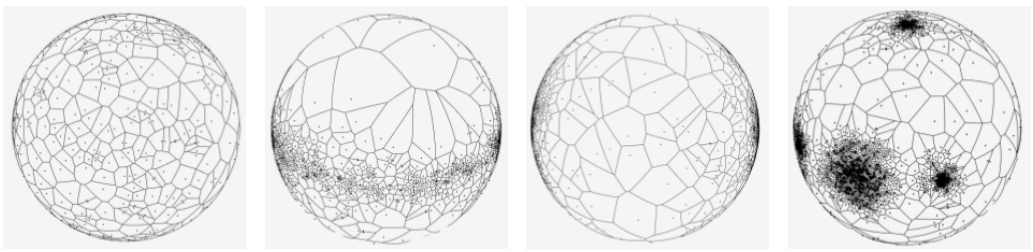


(a) Average per saccade performance plots. $\alpha = 100$



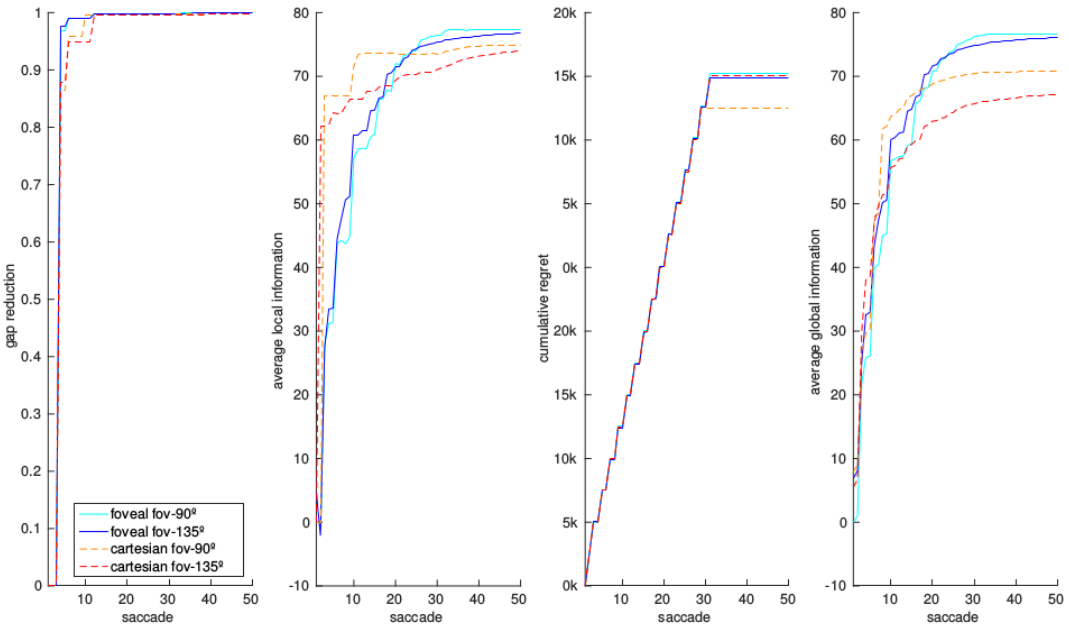
(b) Time evolution plots.

Figure 3.12: Performance results for the assessed memory biases.

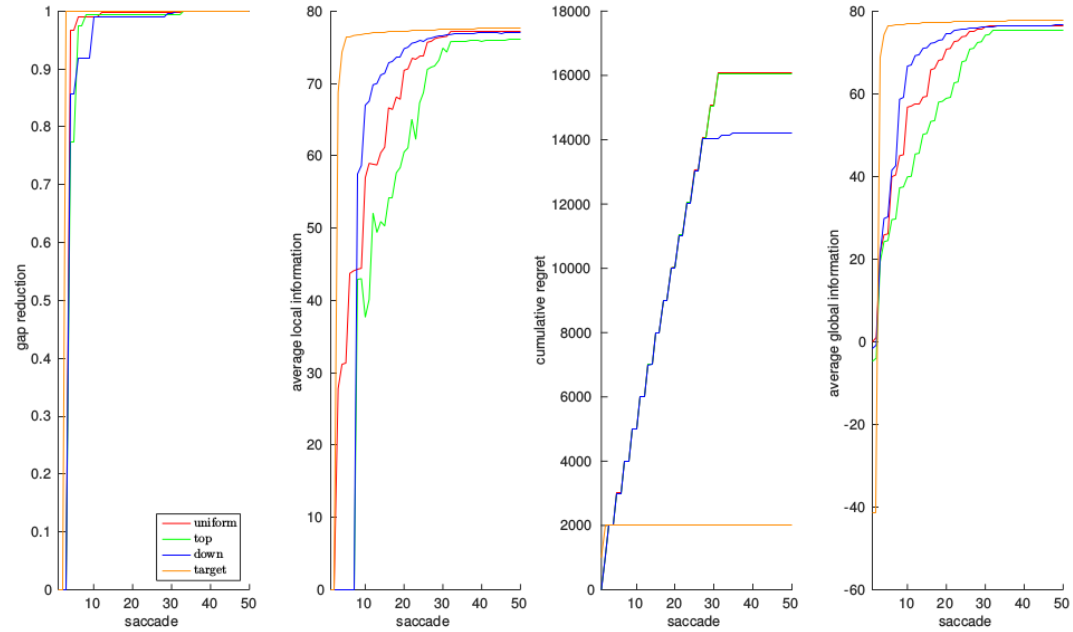


(a) Unbiased. (b) Equator-biased. (c) Anti-podal. (d) Non-trivial.

Figure 3.13: Examples of SES tesselations obtained with the proposed GMM



(a) Time evolution plots for the assessed sensor topologies and field of views.



(b) Time evolution plots for the assessed memory biases.

Figure 3.14: Performance results for the proposed 3D object reconstruction with space-variant binocular and memory mechanisms.

Chapter 4

Pose Estimation with Space-Variant Orientation Selectivity Priors

In this chapter, a complete solution is provided for detecting and identifying parametric shapes, more specifically cylindrical, which are convenient object shapes for studies regarding selectivity to orientations, and which are commonly found in household and industrial environments. Most standard approaches to detect and identify cylinders are not robust to detection of points that lie on the top base, i.e. outliers, which limits their applicability in realistic scenes. In addition, these methods fail to benefit from environmental constraints, e.g. the fact that cylinders often lie or stand on flat surfaces. To tackle the aforementioned limitations, we introduce a novel soft voting scheme that incorporates curvature information in the orientation voting phase. For each potential point on a cylinder, the principal curvature direction is combined with the normal vector to disambiguate candidate orientations. Furthermore, we propose a pre-attentive learning mechanism to selectively discard irrelevant shapes before further processing to avoid time-consuming parametric fitting of wrong detections, thus increasing the efficiency of the whole pipeline. A set of experiments with synthetically generated data are used to assess the robustness of our fitting method with different levels of outliers and noise.

The results demonstrate that incorporating the principal curvature direction within the orientation voting process allows for large improvements on cylinders parameters estimation. Furthermore, we demonstrate that combining bottom-up 3D segmentation with top-down shape-based attention allows for large speed-up and accuracy improvements on cylinder identification. The qualitative and quantitative results with real data acquired from a consumer RGB-D camera, confirm the advantages of the proposed framework.

4.1 Introduction

Due to recent technological advances in the field of 3D sensing, range sensors have become financially affordable to the average consumer, boosting the proliferation of robotics applications requiring accurate 3D object recognition and pose estimation capabilities. More specifically, in tasks that involve interaction with the surrounding environment, an artificial agent would require to accurately recognise objects and estimate their pose. These tasks include successful manipulation and grasping, obstacle avoidance and self localization with respect to known landmarks, to name a few.

Efficiency is another important requirement in robots with power limitations [145], where fast and accurate perception is required, e.g. for the manipulation of kitchenware objects [57]. Therefore, it is of the utmost importance to build efficient perceptual systems that are not only robust to sensory noise, but also to occlusion and outliers. A key aspect behind the success of a grasping solution resides in the choice of the object representation, which can deal with incomplete and noisy perceptual data and is flexible enough to cope with inter and intra-class variability, allowing the generalization to never-seen objects. Furthermore, in order to cope with limited computational processing capacity limitations, efficient and fast perception is an essential requirement for real-time performance. In this work, we propose a computationally efficient attention framework for the task of simultaneously detecting, recognizing and identifying particular object shapes.

We focus on cylindrical shaped objects which are commonly found in domestic (e.g. cups, bottles) and industrial environments (e.g. pipes, pillars, scaffolds), and identifying them plays an important role in many robotic grasp applications [57, 138]. The proposed framework relies on the tabletop assumption, i.e., objects are placed on flat surfaces, which is another widely adopted scenario in robotics [43, 132]. In order to deal with cluttered environments which are often populated with multiple non-cylindrical shapes i.e. distractors, we take advantage of the recent advances in deep learning architectures to introduce an efficient recognition module that learns to filter out irrelevant object candidates. More specifically, we incorporate a pre-attentive shape-based selection mechanism, that avoids the need of time-consuming, top-down cylinder parameter identification at an early stage, on irrelevant salient candidate objects. Furthermore, the most successful cylinder fitting approaches in the 3D shape fitting literature are based on a computationally efficient 2-step Generalized Hough Transform (GHT) [168]. We extend this method with a set of improvements that allow coping with large levels of outliers, mainly residing on bases of cylinders, which often introduce problematic biases during the orientation estimation. The cylinder fitting approach

described in this work was originally proposed in [56], but the reviewed literature and experimental evaluation here is significantly expanded. Also, one of the most successful cylinder fitting approaches in the state-of-the-art [168] is based on a computationally efficient 2-step Generalized Hough Transform (GHT). We extend the former method with a set of improvements that allow coping with large levels of outliers, mainly residing on flat surfaces of cylinders, which often introduce problematic biases during the orientation estimation.

Our main contributions are threefold: first, and unlike previous approaches that are solely based on 3D depth information, we combine a state-of-the-art [168, 56] cylinder fitting approach which is based on a robust and computationally efficient 2-step Generalized Hough Transform (GHT) with a 2D image-based top-down proposal rejection mechanism to increase the quality and speed of correct estimations. Since gathering a large dataset, required for deep learning based recognition techniques is laborious and time consuming, we provide a semi-automatic data gathering procedure, using 3D information, which greatly facilitates acquiring and labeling relatively large amounts of data. Second, we propose a novel randomized sampling scheme for the creation of orientation Hough accumulators.

Our sampling method allows incorporating prior structure knowledge to improve accuracy with fixed computational resources. And finally, as our third contribution, we introduce a novel soft-voting scheme, which considers surface curvature information, in order to cope with points that exist on flat surfaces and that vote for erroneous and arbitrary tangential orientations. We perform a systematic and thorough quantitative assessment of the influence of noise and outliers on detection and pose estimation error of cylinder fitting methods, comparing our proposed method with that of [168]. Our ROS [166] and Caffe [106] C++ implementation can identify multiple cylinders under a second, allowing an easy and straightforward integration in general robotics systems, e.g. in grasping and manipulation pipelines. The code and dataset of our experiments will be released when the final version of this manuscript is prepared. The remainder of this chapter is structured as follows. In section 3.2 we overview previous related work available in the literature. In section 4.3 we describe in detail the various steps involved in the proposed cylinder detection and identification methodology. In section 4.4 we quantitatively evaluate the benefits of the proposed contributions. Finally, in section 4.5 we draw our conclusions and propose promising future work ideas.

4.2 Related Work

As described in the previous section, successful identification of objects in an environment requires not only the development of robust and efficient object detection architectures, but also the definition of flexible shape representations that should facilitate generalization to never-seen-objects, via the integration of different visual sensing modalities. Therefore, we organize the present section in two distinct parts. First, an overview of the state-of-the-art methods in visual attention, with an emphasis on shape-based models of selective attention is presented. Afterward, we analyse various object identification paradigms proposed in the literature, suitable for applications that require identification and localization of parametric shapes.

4.2.1 Shape-based Selective Attention

Visual attention plays a central role in biological and artificial systems to control perceptual resources [6, 158]. The classic artificial visual attention systems use salient features of the image, benefiting from the information provided via hand-crafted filters. Recently, deep neural networks have been developed for recognizing thousands of objects and autonomously generate visual characteristics that are optimized by training with large data sets. Besides their application in object recognition, these features have been very successful in other visual problems such as object segmentation [78], tracking [82] and visual attention [223]. Evidence from neurophysiology studies [65] suggests that people consider oject shape as an important feature dimension among other low-level visual features (e.g. texture and color). In [194] the authors found that subjects looking for a particular shape (e.g. flowers

or pillows) are more accurate in reporting other features of that object (e.g. color) meaning that people have attention mechanisms for shape features. Furthermore, infants rely more on shape than on color when learning new objects, which in turn allows them to generalize to other objects with similar visual features while interacting with them [60]. This fact motivates the need of developing more sophisticated, shape-biased and bottom-up attentional architectures [198].

4.2.2 Object identification in robotics

Object recognition and pose estimation with 3D depth data is an important subject in computer vision with many applications in robotics. There are two main approaches to this problem that depend on the availability of 3D object models: 3D model based and learning based. If one has a description of the 3D shape of the object, either given by a parametric surface representation or by a CAD mesh representation, the 3D model-based methods are often used for simultaneous object recognition and 3D pose estimation [75]. If such representations are not available, the dominant approaches rely on machine learning techniques that "learn a model" given a set of image samples of the object, acquired by the robot sensors [165]. Despite being flexible and capable of generalizing to novel objects in detection and classification tasks, these methods are often unsuitable for estimating some shape properties, such as 3D pose or size of the object. In this work we leverage the accuracy and generalization capabilities of state-of-the-art deep learning techniques in recognition tasks, with robust 3D model-based fitting approaches to develop a multi-modal, fast, and robust cylinder identification pipeline. These representations are often unsuitable, when flexibility and generalization to novel objects is a requirement. The dominant strategies rely in machine learning techniques that are able to generalize to similar objects using a set of sample images acquired by the robot sensors. We focus on cylindrical shapes and, thus, we will combine the generalization capabilities of state-of-the-art deep learning techniques with robust 3D model-based fitting approaches. One of the most successful approaches for model-based 3D object recognition using point clouds are based on [48, 55] where a global descriptor for a given object shape model is created, using point pair features. The CAD model of the object is used to create a large database of features. At run-time, the matching process is done locally using an efficient and robust voting scheme similar to the Generalized Hough Transform [91]. Each point pair detected in the environment casts a vote for a certain object and 3D pose. However in unstructured environments, existing Computer-Aided Design (CAD) based methods tend to suffer from outliers and occlusion. In semi-structured environments (e.g. industrial pipelines), strategies based on the detection and estimation of parametric shapes are generally more robust and flexible [201][150][98]. For the extraction of simple geometric shape primitives like planes, cylinders, cones and spheres, the two most common paradigms are the Hough transform [91] and Random Sample Consensus (RANSAC) [58], which are robust to outliers and noisy data.

RANSAC-based approaches are typically preferred over the former since they are more general and do not require the definition of complex transformations from 3D input to parametric spaces. In the RANSAC paradigm, the data is used directly to compute best-fit models. Despite their proven applicability for the extraction of geometric primitives in noisy 3D data [182] [73], in particular in tabletop object segmentation, RANSAC-based techniques have high memory requirements. Being a non-deterministic iterative algorithm, computational time is greatly dependent on the allowed iterations to produce reasonable results, hence becoming impractical for scenarios with large levels of outliers [125]. In other words, the large number of random selections in large-scale point clouds may compromise the method applicability in applications with real-time constraints. Furthermore, their lack of flexibility hinders the incorporation of model-specific heuristic knowledge, that enables the creation of more effective and efficient specialized methodologies.

The problem of detecting and estimating the pose of cylinder structures using 3D range data and Hough transform is naturally formulated on 5-dimensional parametric spaces, but this results in prohibitive computational complexity due to the curse of dimensionality (the size of the Hough accumulator is exponential in the number of dimensions). A more efficient approach [168] uses a 2D Hough transform to estimate orientation followed by a

3D Hough transform to simultaneously detect radius and position. Though reducing the exponential complexity factor, this approach still lacks speed and robustness in dense point cloud data. In [195] the authors proposed a coarse-to-fine voting procedure that speeds-up the former method by several orders of magnitude [159]. Another interesting idea is the incorporation of environment structural constraints (e.g. cylinders are standing vertically or horizontally on the floor) to reduce the search space [125] to a small subset of possible orientations.

The problem of detecting and estimating the pose of cylinder structures using 3D range data and Hough transform is naturally formulated on 5-dimensional parametric spaces (2 orientations, 2 locations plus the radius), but this results in prohibitive computational complexity due to the curse of dimensionality (the size of the Hough accumulator is exponential in the number of dimensions). The most efficient parametric shape fitting methods are based on Hough transforms that estimate cylinder parameters, i.e. orientation, position and radius, in two sequential voting steps [168, 56]. More specifically, they rely on a 2D Hough transform to estimate orientation, i.e. the direction of the cylinder axis, followed by a 3D Hough transform to simultaneously detect radius and position. Though reducing the exponential complexity factor, this approach still lacks speed in dense point cloud data. In [195] and [159] the authors proposed a coarse-to-fine voting procedure that speeds-up the former method by several orders of magnitude. Another interesting idea is the incorporation of environment structural constraints (e.g. cylinders are standing vertically or horizontally on the floor) to reduce the search space [125] to a small subset of possible orientations.

Despite the improvements on computational complexity of the previous approaches, their lack of robustness to outliers still sets the main draw back to their usage in real applications. Palánz et. al. [156] introduces a method that finds the cylinder that fits better in a point cloud, modeled as a mixture of two Gaussians. One Gaussian models the data samples belonging to the cylinder and the other Gaussian models the outliers. The random variable of the model is the fitting error, which is lower for the inliers and larger for the cylinder outliers. The error considered in their work is the sum of the perpendicular distance from the point to the estimated cylinder, and its parameters are estimated using the Expectation Maximization algorithm for the mixture of Gaussians. Although they show a large robustness to outliers, the method is computationally demanding and not parallelizable. Tran et. al. [201] propose an algorithmic approach that starts from individual cylinder detection, followed by a mean shift clustering in the cylinder space parameters. The individual cylinder detection algorithm finds promising cylinder hypotheses based on weighted point cloud normal estimation and an inlier point selection. The normals are utilized to find the cylinder axis orientation by selecting the eigenvector corresponding to the smallest eigenvalue of the covariance matrix C of normal vectors of inliers. The inliers are selected by projecting the cylinder points to a plane normal to the cylinder axis orientation and fitting the projected points to a circle. This approach is robust to outliers and finds multiple cylinders, but is computationally more expensive than [168], which is the baseline of our approach. Nurunnabi et. al. [150] propose an algorithmic approach that relies on Robust Principal Component Analysis (RPCA) to find the cylinder orientation and Robust Least Trimmed Squares (RTLS) regression to remove outliers from the RPCA cylinder parameter estimation. The RTLS removes outliers that do not fit the projected circle from the cylinder points. This approach is limited to find just one cylinder in the point cloud.

In this work we propose a novel fitting approach that leverages an efficient implementation of the Hough-based method of [168] with the increased robustness of using statistical models to encode domain-specific knowledge. More specifically, the focus and the main contributions of our work are: a novel randomized sampling scheme for the creation of orientation Hough accumulators which allows the incorporation of environment structural priors to improve orientation estimation accuracy with the same computational resources; a voting scheme that significantly improves the robustness of Hough methods in cylinder detection and pose estimation.

Still, all the aforementioned fitting approaches are incapable of filtering, at an early stage, different object shapes that act as irrelevant visual distractors. The time consuming process of fitting shapes to distractors, marks another limitation of fitting approaches, which hinders their applicability in real world scenarios.

Kostavelis et al. [113] have incorporated Graph-Based Visual Saliency algorithm (GBVS) as a pre-processing step in training a biologically inspired Hierarchical Temporal Memory (HTM) network. According to these results,

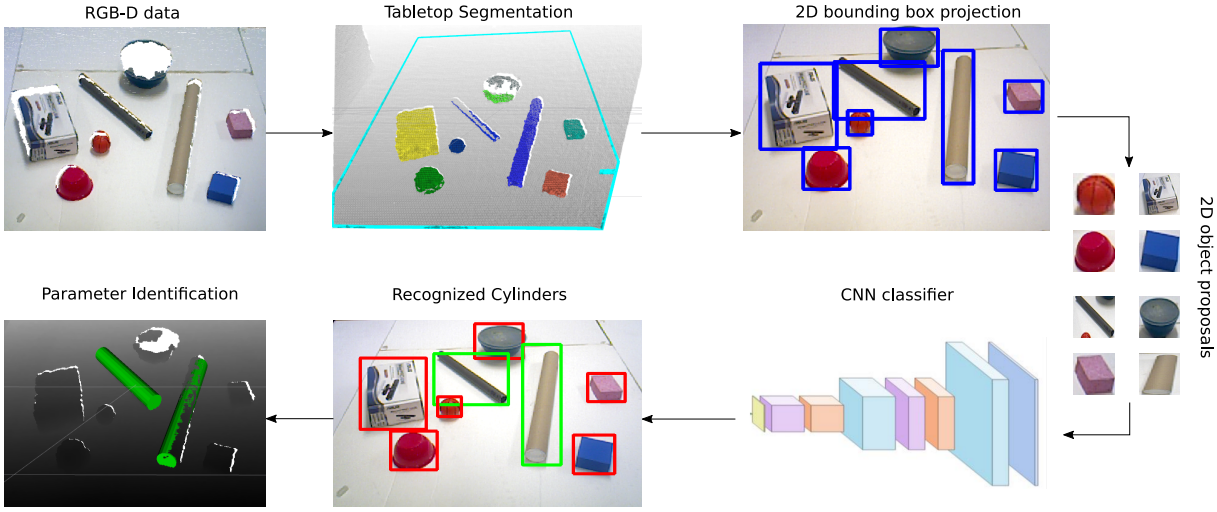


Figure 4.1: General diagram describing our framework for efficient detection and identification of cylindrical shapes using multiple visual sensing modalities: color and depth. The proposed architecture, is an integration of different cognitive blocks which are responsible for object segmentation, shape recognition, and fitting.

the introduction of a bottom-up attention mechanism significantly improves the efficiency and performance of down-stream tasks, however, it is not clear how much their approach can generalise to the detection of occluded objects. Similarly, we incorporate a mediating shape-based pre-attention bottom-up mechanism to reduce the space of possible cylindrical shapes to a small subset of prominent objects in the field of view, in a bottom-up manner. The 2D image patches, coming from 3D segmentation are first classified using a Deep Convolutional Neural Network (DCNN), which is robust to occlusion. Object classes of interest (i.e. cylinder), are further considered for parameter identification, which results in faster and more accurate estimates.

4.3 Methodologies

In this section we describe in detail the multiple components and contributions of our pipeline (see Figure 4.1).

4.3.1 System Overview

We start by detecting tabletop objects using 3D point cloud information, since points above tables are considered to belong to potentially graspable objects. Therefore, the first component of our cylinder detection and identification pipeline is a bottom-up segmentation module that is triggered by salient objects laying on flat surfaces [146]. First, we use a RANSAC-based fitting approach, which efficiently operates on downsampled organized point cloud data [180], in order to detect planes on the scene and segment objects above these planes. We rely on Euclidean clustering [180] to identify individual objects. Afterwards, these objects are projected on the 2D camera plane to extract bounding boxed 2D focused images from a stream of monocular images, which are used to recognize cylindrical shapes via a deep artificial neuronal network classifier. The proposed CNN is trained offline via transfer learning, and acts as a shape-based mediating pre-attentive selective mechanism that filters out non-cylindrical shapes. Finally, the parameters of the identified cylindrical shapes are estimated in 3D Cartesian space, using an efficient and robust top-down depth-based Hough transform.

4.3.2 Transfer learning for early shape-based attention

In order to reject region proposals and avoid parametric identification of non-cylindrical objects, we propose to use deep neural networks. Inspired by recent advances of deep learning in achieving state of the art performance in recognition tasks, we use a deep CNN as a binary classifier to decide if a particular object is a cylinder or

not. However, using a DNN for the task at hand can pose several challenges. Firstly, most DNN architectures are notoriously data-hungry, usually trained on millions of labeled images. Secondly, designing a neural network architecture for a new task is time consuming and involves a large amount of trial and errors. And last, storing and using them on most embedded systems are impractical due to the substantial size and the computations they require.

Data acquisition and training

To solve the first problem, we propose a fast and convenient procedure for semi-automatic gathering of labeled data, which does away with the need of manual labeling. The procedure relies on the 3D tabletop segmentation method and the 3D bounding box projection to 2D approach described in the previous subsection. For the creation of positive samples, we first place many different cylindrical shaped objects on tabletops and acquire data, from multiple views, using an hand-held RGB-D camera. Then for the creation of the negative examples dataset, we repeat the same procedure with all the non-cylindrical objects, commonly found in the testing environment.

Cylindrical-shapes recognition

For the second problem, i.e. architecture design, we propose to use transfer learning [217]. More specifically, we have used a network previously trained on imagenet dataset [179] and fine-tuned it as a cylinder classifier. This way, the architecture of the network is pre-defined and it is only necessary to change the last layer such that instead of predicting probability classes of 1000 objects, it only outputs the probability that an input image is a cylinder or not. Moreover, it is generally assumed that if a network performs well on a recognition task, it means it has learned *informative* features which are useful for different tasks. As a result, it is possible to train the network on significantly smaller datasets and only slightly change the previously learned features.

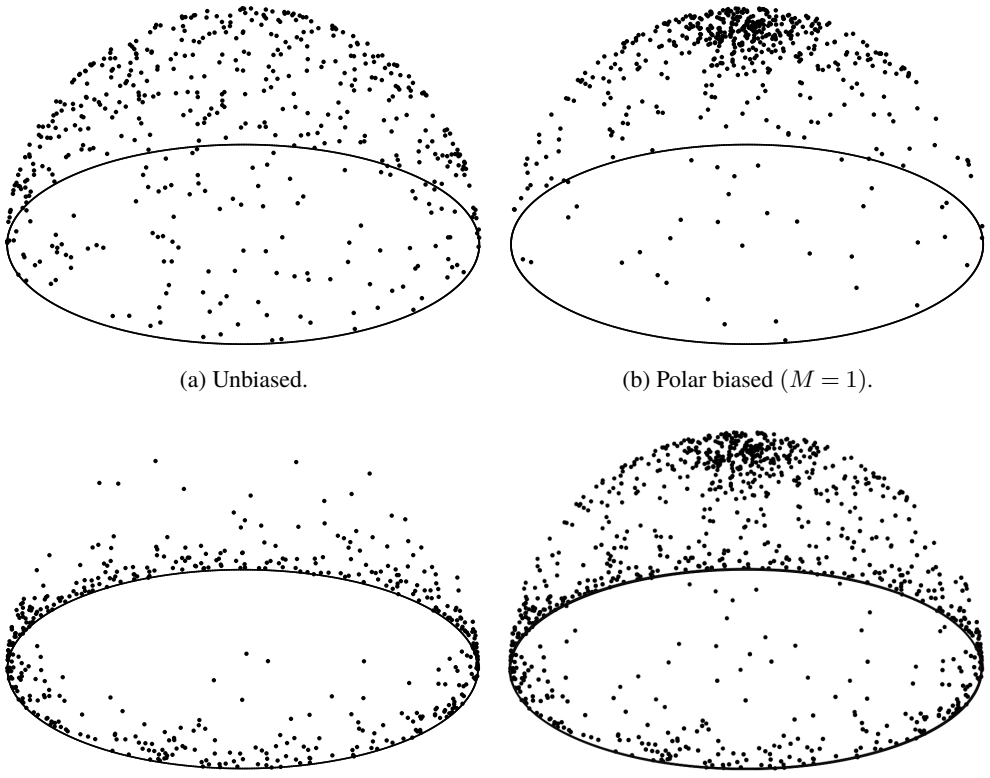
Performance speed-ups

In order to have a small network which performs reasonably fast even in the absence of powerful GPUs, we used a relatively small neural network called SqueezeNet [99]. This network achieves AlexNet accuracy score on imagenet while being 50 times smaller. Taking advantage of this reduction in parameters of the network, it is possible to have a fast and reliable classifier which is more suited towards real-time applications. However, although we have chosen Squeezenet as our object classifier since, at the time, it was the one that offered the best trade-off between performance and efficiency, our method is flexible enough to easily incorporate any other object classifier.

4.3.3 Cylinder parametric fitting

Our approach is based on the former work of Rabbani et al. [168] that splits the cylinder detection and pose estimation problem in two independent Hough transform stages. In the first stage, 3D point normals cast votes for possible cylinder orientations, in a 2D orientation accumulator. In the second stage, the point cloud is rotated according to the determined orientation and each point votes for a position and radius of the cylinder in a 3D Hough accumulator. In that work the unit sphere of orientations is uniformly and deterministically sampled at a predefined number of points [130], to generate a discrete Hough accumulator space, in which voting is subsequently performed. A larger number of cells on the unit sphere improves the accuracy of the orientation estimate, at the cost of increased computational effort. In the present work, we propose several improvements to the orientation voting stage of [168].

In this section we describe in detail our methodology for improved orientation estimation during cylinder detection. First, we introduce a novel randomized sampling scheme which enables the creation of non-uniform, problem-specific orientation Hough accumulators. Then we present a novel and more efficient Hough voting scheme that relies on simple inner products. As opposed to [168], we avoid the computational burden of explicitly



(a) Unbiased.

(b) Polar biased ($M = 1$).

(c) Equator biased ($M = 1$)

(d) Non-trivial ($M \geq 1$)

Figure 4.2: Different sampled unitary spheres, where each point on the unit sphere represents the center of a candidate Voronoi cell orientation.

voting in spherical coordinates, which requires the computation of rotation matrices and, consequently, of inefficient trigonometric functions. Furthermore, our voting scheme is richer than the one of [168] since it allows incorporating curvature information. When compared with the work of [168], the proposed methodology is able to cope with higher levels of outliers, including flat surfaces such as ground planes, hence avoiding the need of prior plane detection and removal.

Randomized Orientation Hough Accumulator

The proposed orientation Hough accumulator space is composed of a set of cells $\mathcal{D}$ lying on a unit sphere. The center of each cell corresponds to a unique absolute orientation. The accumulator is analogous to a Voronoi diagram defined on a spherical 2-manifold $\mathbb{S}^2$ in 3D space, as depicted in Figure 4.2, and is represented by a set of N_d 3D Cartesian sample points with unit norm, centered in the reference frame origin (center of the sphere) $\mathcal{E}$,

$$\mathcal{D} = \{\mathbf{d}^i \in \mathbb{R}^3, i, ..., N_d : \|\mathbf{d}^i\| = 1\} \tag{4.1}$$

which are i.i.d. and randomly generated from a three dimensional GMM distribution

$$\mathbf{d}^i = \frac{\mathbf{v}^i}{\|\mathbf{v}^i\|} \text{ where } \mathbf{v}^i \sim p(\boldsymbol{\theta}) = \sum_{m=1}^{M} \phi^m \mathcal{N}(\boldsymbol{\mu}_d^m, \boldsymbol{\Sigma}_d^m) \tag{4.2}$$

where M is the number of mixture components and where each $\mathbf{d}^i \in \mathcal{D}$ represents an absolute orientation, allowing for efficient voting with observed surface normals.

The parameters of the GMM components are chosen according to task at hand (e.g. find vertically aligned cylinders) or prior knowledge on how likely specific orientations are (e.g. cylinders are unlikely to be in relative diagonal orientations). On one hand, in order to produce uniform and unbiased accumulator structures, the surface

should be sampled from a rotationally symmetric distribution, i.e., from a single Gaussian with zero mean and variance equal in all dimensions [147] (Figure 4.2a). On the other hand, non-uniform, task-dependent sampling biasing can be achieved by manipulating the Gaussian Mixture Model parameters (see Figure 4.2).

Hypothetical accumulator spaces that may be suitable for different priors are depicted in Figure 4.2. In the absence of prior information or task definition, one should sample from a single component Gaussian, with zero mean and standard deviation equal in all dimensions (Figure 4.2a). If for instance the task is to find cylinders that are vertically aligned with the reference frame (e.g. table reference plane), one should privilege orientations at the pole (Figure 4.2b) rather than the equator (Figure 4.2c). In the latter case, varying the Gaussian mean is not sufficient. One could sample from a single-component zero mean GMM with larger variance in the horizontal directions. Finally, prior knowledge or more complex detection tasks (e.g. locating diagonal pipes or machine handles) can benefit from a GMM with many components (Figure 4.2d).

Our randomized sampling scheme offers several advantages over the one of [168], namely:

- it is easier to implement than its deterministic counterpart [130] and allows for the fast creation of biased orientation voting spaces.
- the non-deterministic nature of the representation offers a convenient mechanism for encoding task-related biases or probabilistic prior knowledge about possible orientations, depending on the environment (e.g. cups are typically oriented vertically on tables). Biasing the orientation Hough accumulator space leads to more efficient, flexible and adaptable resource allocation and to more accurate orientation estimation, for the same memory and computational resources.

Fast Robust Orientation Voting Scheme

At run-time time, the input of our algorithm is a scene input point cloud which comprises a finite set of 3D Cartesian points $\mathcal{P} \subset \mathbb{R}^3$, where $P = \{\mathbf{p}^s, s = 1, ..., N_s\}$.

First, we estimate the surface normals at each scene point $\mathbf{p}^s \in \mathcal{P}$ using the Principal Component Analysis (PCA) [52] of the covariance matrix created from its k-nearest neighbors. Let $\mathcal{N} = \{\mathbf{n}^s, s = 1, ..., N_s\}$ denote the set of surface normals. Then, we proceed with the computation of the principal curvatures as follows. For each scene point $\mathbf{p}^s$, we compute a projection matrix for the tangent plane given by the associated normal $\mathbf{n}^s$. After, we project all normals from the k-neighborhood onto the tangent plane. Finally, we compute the centroid and covariance matrix in the projected space. We finally employ eigenvalue decomposition of this covariance matrix to obtain the principal curvature direction $\mathbf{c}^s_{\text{max}} \in \mathbb{R}^3$ and the corresponding eigenvalue $k_{\text{max}} \in \mathbb{R}$ (see Figure 4.3). Let $\mathcal{C} = \{\mathbf{c}^s_{\text{max}}, s = 1, ..., N_s\}$ denote the set of principal curvature directions and $\mathcal{K} = \{k^s_{\text{max}}, s = 1, ..., N_s\}$ the set of the corresponding eigenvalues.

The orientation voting procedure goes as follows: For each direction cell $\mathbf{d}^i$ in the orientation Hough accumulator A, we compute the inner product with all the scene surface normals $\mathbf{n}^s \in \mathcal{N}$ and their associated principal curvature directions $\mathbf{c}^s_{\text{max}} \in \mathcal{C}$ to cast continuous votes in the accumulator according to the function

$$A(i) = \sum_{s=1}^{N_s} k^s_{\text{max}} \left|\left(1 - \mathbf{d}^i \mathbf{c}^s_{\text{max}}\right)\right| \left|\left(1 - \mathbf{d}^i \mathbf{n}^s\right)\right| \tag{4.3}$$

This soft voting function gives more weight to directions that are simultaneously, orthogonal to the normal and the principal curvature directions. Moreover, the eigenvalue k^s_{max} works as a curvature high-pass filter, that suppresses low curvature candidates, since points belonging to flat surfaces have very low k^s_{max}.

After determining the cylinder orientation we proceed with the estimation of the cylinder position and radius, as detailed in [168]. First, we align the estimated cylinder axis with the camera z-axis. Then, we project the inlier points on the camera xy plane and use a Circular Hough Transform (CHT) [110] to estimate the cylinder position and radius.

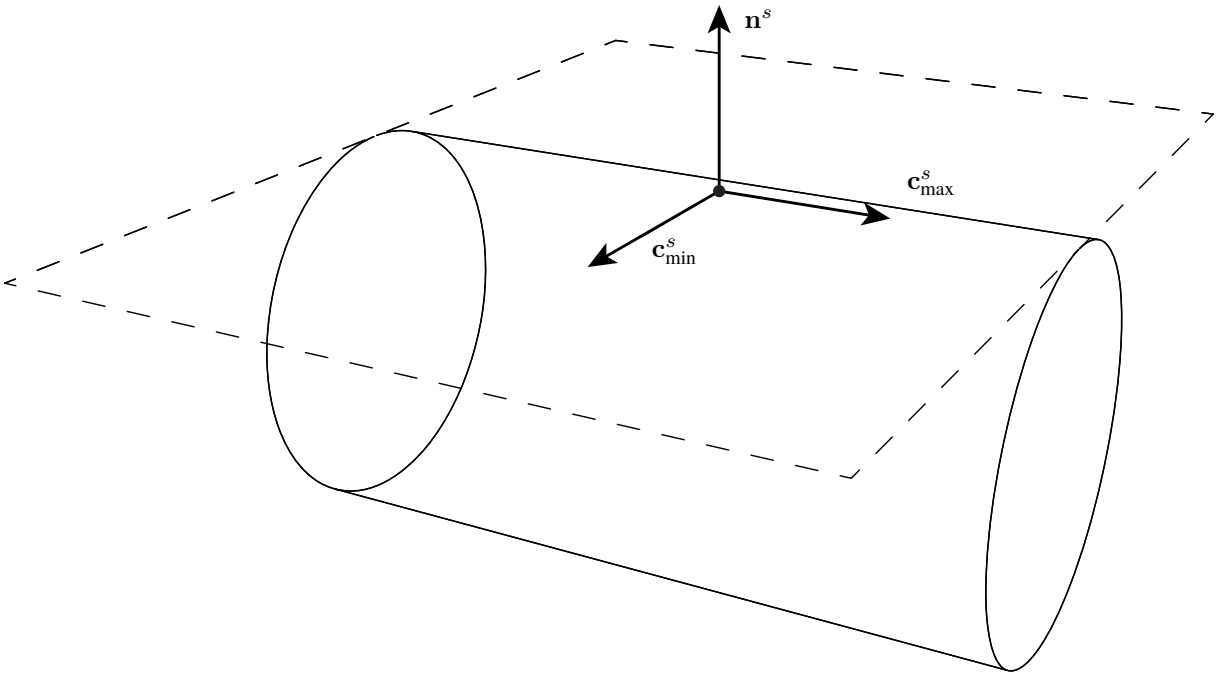


Figure 4.3: Normal ($\mathbf{n}^s$) and principal curvatures' directions ($\mathbf{c}^s_{\text{max}}$ and $\mathbf{c}^s_{\text{min}}$) for a cylinder surface point.

Goodness-of-fitting criterion

Finally, the goodness of the fitting of a cylinder is evaluated using the following conditional confidence measure:

$$p(\text{cylinder}|\text{object}) = \frac{N_{model}}{N_{\text{cluster}}} \tag{4.4}$$

where N_{model} represents the number of points that fit the estimated cylinder parametric model (i.e. inliers) and N_{cluster} the total number of 3D points belonging to the object. Estimations below a user-defined quality threshold are discarded and considered as non-cylindrical shapes. We have used this criterion as a *baseline* for cylinder detection.

4.4 Experiments and Results

Several experiments were conducted in order to quantitatively evaluate the quality of the cylinder parameters recovered by the method of Rabbani et al. [168] and by our proposed method, when dealing with increasing levels of outliers and noise.

4.4.1 Synthetic Data

In all experiments, we generated 200 synthetic scenes, each containing a single instance of a cylinder. By using synthetically generated scenes, we were able to compare the algorithm pose results with a known ground truth. The selected parameters for both methodologies were the following: The radius was fixed to $r = 0.3m$ and the height was uniformly sampled from the interval $[0.05, 2.0]\, m$. The number of cylinder surface points was fixed and set to 900 and the number of orientation sample points in the Hough accumulator space was set to $N_d = 450$.

In order to demonstrate the advantages of incorporating prior knowledge in the creation of orientation Hough accumulators, in all generated scenes the orientation of the cylinder was fixed and aligned with the z-axis of the frame of reference, ignoring other directions, e.g. horizontal, to also facilitate the estimation error statistics (i.e. averages and standard deviations). We considered and compared the following different sampling distributions for creating the orientation Hough accumulator space (see Table 4.1):

- an unbiased distribution reflecting the absence of prior knowledge about the cylinder orientation.

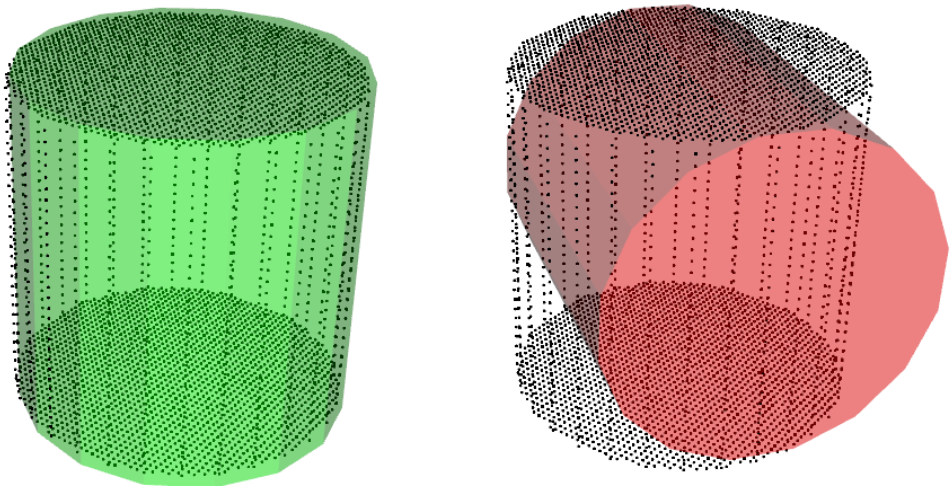
- a mildly and a strongly biased distribution that favour vertical orientations.

Robustness to outliers

In order to assess the performance gains of the proposed strategies in the presence of flat surfaces (i.e. outliers)

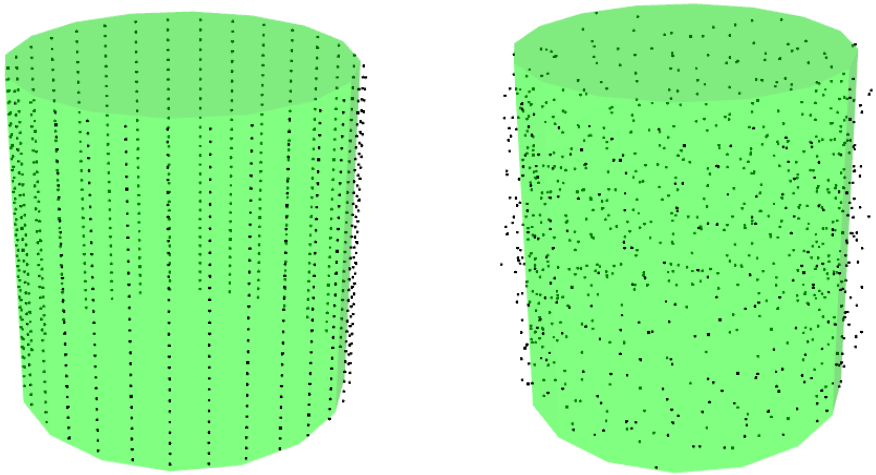
$$\text{outliers} = 1 - \frac{\text{cylinder surface points}}{\text{total scene points}} \tag{4.5}$$

we added synthetically generated planar extremities to cylinders, that simulate realistic cylindrical shapes such as containers/cans with lids. Surface points on cylinder tops are problematic for orientation estimation since they vote for orthogonal directions, and in this experiment were considered as planar clutter (i.e. statistical outliers). The surfaces were generated with a total of 10, increasing point density levels, to each previously generated cylinders' bottom and top extremities (see Figure 4.4). Moreover, each cylinder was set at a random pose. The quantitative results illustrated in Figure 4.9 (left column) demonstrate the advantage of considering both the surface curvature and the surface normal in the orientation voting step. When dealing with flat surfaces that belong to real-life cylinders, our method estimates better the cylinder orientation, as shown by the absolute orientation errors in Figure 4.9a and Figure 4.9b. According to our implementation, the original method of Rabbani et al. can deal with cases where up to 50% of the points are outliers, without failing. When the number of outliers exceed 150% of the relative number of candidate points belonging to the cylinder surface, the method exhibits an orientation error of 90 degrees, since points belonging to flat surfaces (i.e. outliers) vote for orthogonal directions to the ground truth cylinder orientation. The linear transition in between can be justified by the fact that the error increases linearly with the number of outliers voting for orthogonal, wrong orientations. This is an artifact of the soft-voting scheme, resulting in consistent response to small and large amount of outliers. In between, the response exhibits a linear decrease in the pose estimation accuracy. As expected, these improvements have a direct and positive impact in the quality of the position and radius estimations, depicted through the absolute radius and position errors plots in Figure 4.9d and Figure 4.9c. As opposed to [168], our method is able to cope with large amounts of outliers, while keeping the performance at the levels of uncluttered scenes.



(a) Ours (b) Rabanni et al.

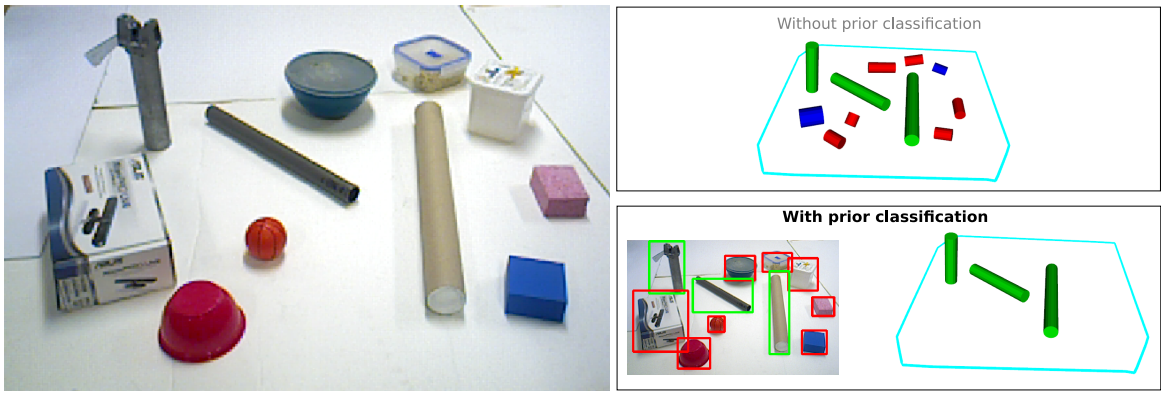
Figure 4.4: Our method against Rabbani et al. when dealing with flat surfaces.



(a) $\sigma_{\text{noise}} = 1\%$ (b) $\sigma_{\text{noise}} = 10\%$

Figure 4.5: Estimated cylinder parameters with our method, from a point cloud corrupted with different levels of additive Gaussian noise.

(a)



(b)

Figure 4.6: Qualitative assessment of our framework with data acquired with an Asus Xtion 3D camera. (a) Testing scene samples. (b) Cylinder identification for an example scene from the collected 200 frame dataset. Detection: Good and bad classifications in green and red, respectively. Parameter identification: green represents correct parameter estimation; blue represents correct non-cylindrical shape objects identified by the baseline quality of fitting criterion; red represents wrong estimations without the classifier.

Bias	$\boldsymbol{\mu}_p$			$\boldsymbol{\Sigma}_p$		
	x	y	z	xx	yy	zz
Unbiased	0	0	0	0.5	0.5	0.5
Mildly Top-biased	0	0	1.0	0.5	0.5	0.5
Strongly Top-biased	0	0	1.0	0.05	0.05	0.05

Table 4.1: Orientation Hough accumulator biasing parameters used for the creation of the orientation Hough accumulators in the experiments with synthetic data.

Figure 4.7: Sample examples from the training dataset after rotation augmentation. (a) Cylindrical samples (b) Non-cylindrical samples.

Robustness to noise

In pursuance of quantifying the behavior of the Rabbani el al. algorithm [168] and our proposed extensions in the presence of noisy visual sensors, each of the 200 generated scenes was corrupted by 10 different levels of additive Gaussian noise, with standard deviation proportional to the cylinder radius (see Figure 4.5). Figure 4.9 (right column) depicts the cylinder parameters estimation errors for both methodologies in the presence of noise. The results show that both methodologies have similar robustness to noise, hence, demonstrating the benefit of our approach when considering the superior performance of our method in cluttered scenes. Additionally, biasing the orientation accumulator in the face of prior structural knowledge significantly improves the estimation accuracy. Overall, our extensions result in dramatic improvements regarding robustness to clutter, without sacrificing robustness to noise. Furthermore, a simple qualitatively assessment of our method with data acquired from a RGB-D camera demonstrates its applicability to real-scenarios, as exemplified in Figure 4.6, and its superior robustness to planar clutter.

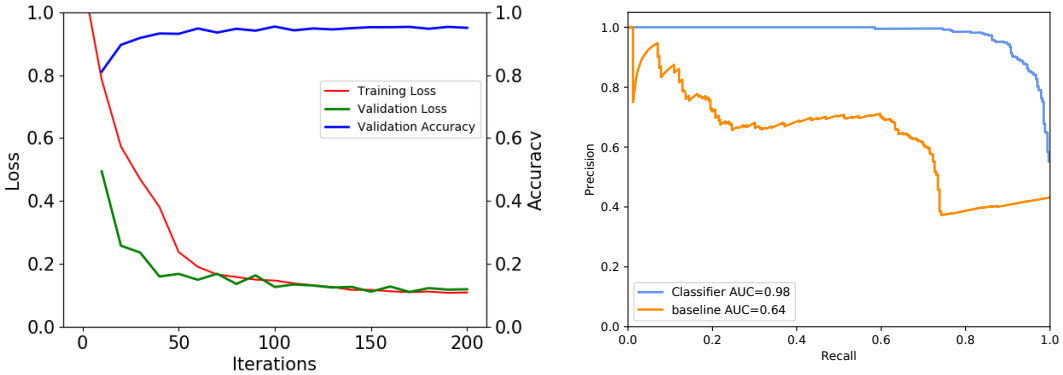
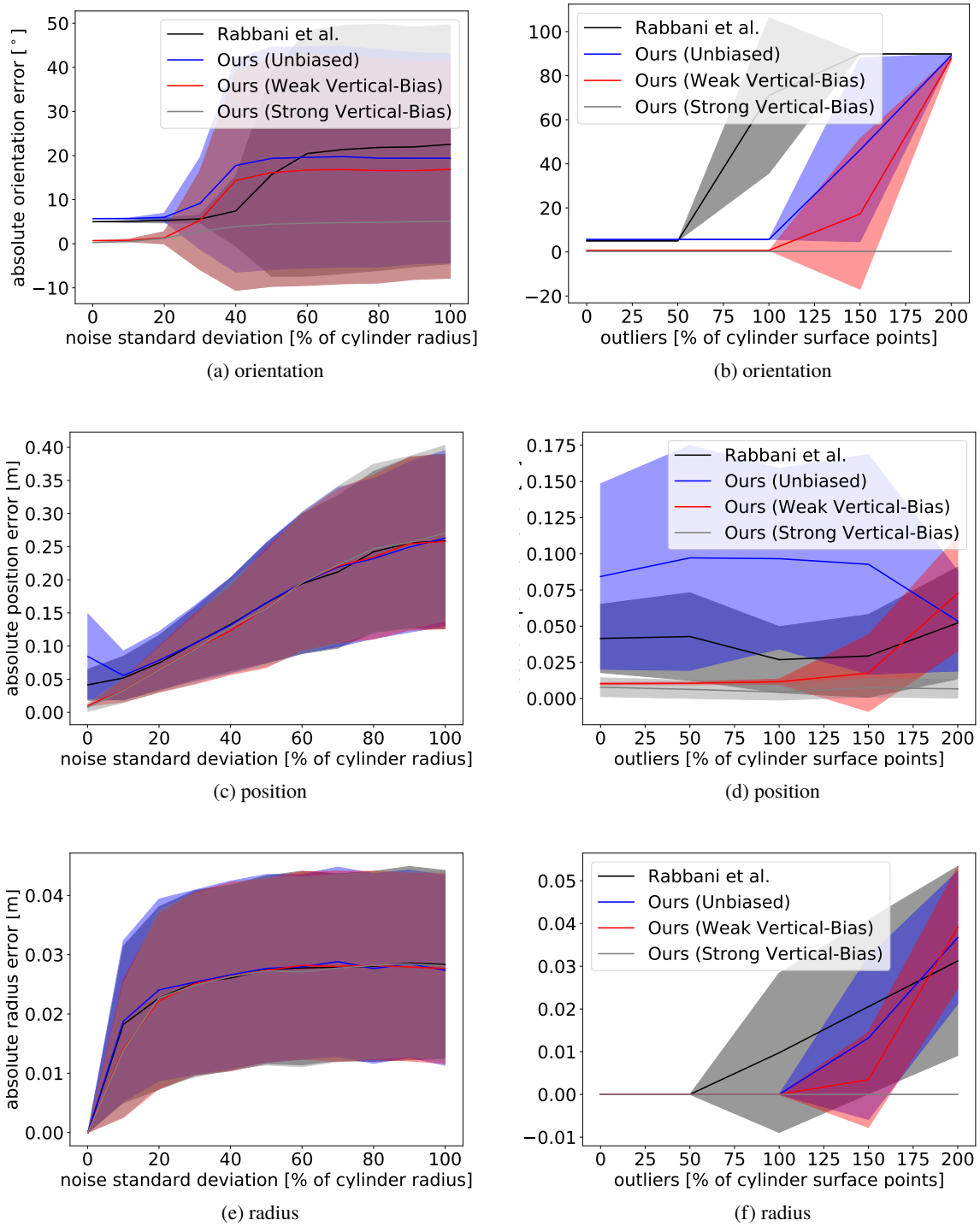


Figure 4.8: Evaluation of the performance of the binary classifier: (a) Loss and accuracy evolution of the classifier on training and validation data. (b) Precision-Recall curves of the Cylinder class for baseline and SqueezeNet classifier on the test data, AUC: Area Under Curve.



(a) orientation (b) orientation

(c) position (d) position

(e) radius (f) radius

Figure 4.9: Robustness of our method against the method of Rabbani et al. Left: different levels of noise. Right: different levels of flat surface outliers.

	Avg. Objects Number		Avg. Processing Times (ms)			
	Cylinders	Distractors	Segmentation	Classification	Identification	Total
F-RCNN	3	8	402		73	475
no classifier	3	8	15	-	213	228
with classifier	3	8	15	64	70	**149**

Table 4.2: Quantitative analysis of the time performance of the proposed pipeline in a set of multiple tabletop scenarios, with 200 RGB-D frames acquired with an Asus Xtion camera.

4.4.2 Real Data

In order to assess the behavior of the proposed framework with real data acquired from a low-cost consumer RGB-D sensing device, we created multiple tabletop scenarios, each containing various different shapes including cylindrical objects (see Figure 4.6a for an example view). We quantitatively and qualitatively evaluated our attentional framework's computational time improvements, in the presence of salient visual distractors.

Classifier Performance Analysis

As described in the previous section, we fine-tune the final layer of SqueezeNet with our newly gathered dataset which contains about 11000 train images (out of which, we used 10% for validation) and 1200 test images. Figure 4.7 shows a few samples that were used to train the network. The original dataset contained less than 3000 samples and, in order to gain more robustness to different orientations, they were mirrored in vertical and horizontal directions, effectively quadrupling the amount of available data. The learning rate for fine-tuning the network was empirically selected as 0.0005 and we kept other parameters as their proposed values by [99]. Figure 4.8 shows the performance of the classifier at various points during training. Our initial experiments with the neural network classifier suggests a generalization to unseen cylindrical and non-cylindrical objects. However, not surprisingly, it is more reliable in classifying seen cylinders. Introducing more unique cylinders can help mitigating this effect. In order to quantitatively evaluate the performance of the 2D image-based deep neural network classifier, it is compared with a baseline indicator of the fit quality criteria. Figure 4.8 compares the precision–recall curves of the two classifiers.

4.4.3 Overall Framework Assessment

In order to qualitative evaluate our framework when dealing with real data we acquired 200 RGB-D frames with resolution 640x480 provided by an Asus Xtion camera. Figure 4.6 depicts the cylinder parameters estimation quality for the proposed cylinder fitting methodology in the presence of noisy 3D point cloud data. The use of prior classification, results not only in temporal gains (see Table 4.2), but also on early filtering of non-cylindrical distractors, hence improving the reliability of the 3D cylinder fitting approach. Overall, dramatic improvements on detection speed and robustness to visual distractors can be achieved by incorporating the of shape-based pre-attention mechanism, results in dramatic improvements on detection speed and robustness to visual distractors without sacrificing robustness to noise. Furthermore, the evaluation of our method with data acquired from a consumer RGB-D camera demonstrates our method applicability to real-scenarios and its advantages in scenes populated with salient visual distractors. In order to better ground the time complexity of this pipeline, we have also experimented with an off-the-shelf state-of-the-art object detector (F-RCNN) [173], which similar to SqueezeNet was also fine-tuned to detect cylinders in RGB images [173]. This detector uses ResNet50 as the classifier and we have reduced the number of region proposals to decrease the inference time. Using the detector, one can achieve a constant run-time with respect to the number of objects in a scene, however, according to Table 4.2, the proposed pipeline is more then twice times faster even with an average of 8 visible objects. Furthermore, unlike off-the-shelf object detectors, 3D tabletop segmentation allows the definition of a table coordinate frame and, hence, the incorporation of prior knowledge in the fitting process.

4.5 Conclusions

In this work, we proposed a complete, robust and efficient cylinder detection and parameter identification framework, that leverages prior knowledge regarding the likelihood of certain object orientations, to improve accuracy with constrained computational resources. Furthermore, and unlike previous approaches that are solely based on 3D depth information, our methodology incorporates RGB information by means of a novel shape-based pre-attentive

attentional mechanism to filter out visual distractors at an early stage. Finally, we have developed a robust soft-voting scheme based on the GHT [15] for the detection and pose estimation of arbitrary cylindrical structures from 3D point clouds. In this work we focused on cylindrical shapes, since they are convenient for studies involving orientation selectivity mechanisms, but the same principles can be easily extended to other types of shapes.

The proposed method incorporates curvature information in the voting scheme, that improves the rejection of outliers, mainly those arising from planar surfaces that pollute the orientation voting space and introduce erroneous biases in cylinder orientation estimation. The results demonstrate significant detection accuracy and time speed-ups as well as major improvements on the detection rates and pose estimations with respect to previous schemes. A systematic quantitative analysis of robustness to clutter and noise validates our approach and sets a benchmark for future research.

The proposed contributions can be summarized as follows:

1. we propose a novel randomized sampling scheme, that allows incorporating allocentric or object-centric structural priors in the creation of orientation Hough accumulators using a GMM. The incorporation of these priors allow improving estimation accuracy with fixed resources, i.e., fixed number of cells in the parameteric orientation acummulator space **P.IV**.

2. we introduce a novel soft-voting scheme, which considers surface curvature information to cope and increase robustness to flat surfaces that vote for erroneous and arbitrary tangential orientations **P.IV**.

3. we take advantage of recent advances on deep learning architectures to introduce an efficient image-based recognition module which learns in a supervised and data-efficient manner to selectively discard irrelevant shapes before further processing 1. We used a relatively small neural network called SqueezeNet [99]. This network achieves AlexNet [118] accuracy score on imageNet while being 50 times smaller. Taking advantage of this reduction in parameters of the network, it is possible to have a fast and reliable classifier.

4. A set of experiments with synthetically generated data were used to compare the robustness of our method, for different levels of outliers, noise and missing data, against the one of [168]. The results demonstrated that the proposed randomized sampling approach for creating Hough orientation accumulators, as well as the incorporation of the principal curvature directions within the orientation voting procedure allows for large improvements on cylinders' parameters estimation. Qualitative results with point clouds acquired from consumer RGB-D cameras, confirmed the advantages of using a cylinder CNN classifier prior to fitting, both in terms of speed and accuracy **P.VI**.

Future work The limitations of the current work include being agnostic to information integration across time (via recursive Bayesian filtering), the generalization of the shape detection method and extension of the Hough orientation estimation to other shape types. The proposed methods for object pose estimation can be improved as follows:

- *Sequential Bayesian Filtering:* robustness to noise can be enhanced via temporal integration of cylinder detections by means of sequential Bayesian filtering [54] techniques.

- *Generalizing to Multiple Shapes:* The idea of combining a generic multi-label classifier with the proposed randomized Hough accumulator and the soft voting scheme, can be extended to other parametric shape types (e.g. cuboids, ellipsoids, cones).

- *Dynamic coarse-to-fine orientation estimation:* Currently, our method relies on single GMM priors, whose parameters are provided before run-time. Pose estimation accuracy would improve if multiple GMM with increasing resolution, and decreased dispersion were considered, and the voting procedure carried out in a coarse-to-fine manner.

We have focused on cylindrical shapes but the proposed core ideas can be easily extended to other shape types, depending on training data availability. Combining a generic multi-label classifier with the proposed randomized Hough accumulator and the soft voting scheme, paves the way to extend the current cylinder identification pipeline to various shapes (e.g. cuboids, ellipsoids, cones). As a final remark, we emphasise that the computational complexity of the proposed solution scales linearly with the number of objects in the scene, which may become problematic in highly cluttered environments. However, all components of the pipeline are parallelizable and, depending on the application requirements, one can benefit from an increase in the available hardware resources to further improve run-time performance. Finally, complex objects such as cylindrical containers require more elaborate representations such as semantic or relational. In the case of cylindrical containers one can consider that containers have two object primitives: planes and cylinders. Future work should consider these type of representations through the use of Probabilistic Graphical Models [81] to further improve the pipeline performance.

Chapter 5

Multiple Object Tracking with Resource Constraints

In this chapter we address the multiple person tracking problem with resource constraints, which plays a fundamental role in the deployment of efficient mobile robots for real-time applications involved in Human Robot Interaction (HRI).

5.1 Introduction

We pose the multiple target tracking as a selective attention problem in which the perceptual agent tries to optimize the overall expected tracking accuracy. More specifically, we propose a resource-constrained POMDP formulation that allows for real-time on-line planning. Using a transition model, we predict the true state from the current belief for a finite-horizon, and take actions to maximize future expected belief-dependent rewards. These rewards are based on the anticipated observation qualities, which are provided by an observation model that accounts for detection errors due to the discrete nature of a state-of-the-art pedestrian detector. Finally, a Monte Carlo Tree Search (MCTS) method is employed to solve the planning problem in real-time. The experiments show that directing the attentional focci to relevant image sub-regions allows for large detection speed-ups and improvements on tracking precision.

The remainder of this chapter is structured as follows. In section 5.2 we overview some related work available in the literature. In section 5.3 we describe the various components involved in the proposed adaptive tracking pipeline. In section 5.4 we assess the proposed methodology performance by evaluating the balance between efficiency (low computational requirements) and effectiveness in multiple object tracking task-execution. Finally, in section 5.5 we wrap up with some conclusions and future work.

5.2 Related Work

MOT deals with the challenging computer vision problem of jointly estimating object identities and motion trajectories in video sequences and has a variety of uses in various application domains. Developing efficient adaptive MOT systems that are capable of dealing with computational and power limitations as well as timing requirements is of the utmost importance in a wide range of fields, including automatic surveillance [192], sports analysis [212] and HRI [137].

Tracking can be framed within different paradigms, depending on the tracklets initialization (*detection-based* vs *detection-free*), and processing mode (*online* vs *offline*) [129].

Detection-based or *tracking-by-detection* approaches assume that candidate detections are provided at each time instant, and the goal of the tracking methodology is to associate detections over time. Detection-free methods, on the other hand, assume tracklets are manually initialized and localization performed in each subsequent frame, in a top-down manner. The former approaches can easily handle appearing and disapearing targets, but require pre-training an object classifier for task-specific objects. As opposed, the latter does not require *a priori* knowledge regarding the objects to be tracked, but require learning online the targets' appearance.

Online tracking approaches employ recursive inference techniques, for sequentially arriving images, while *offline* techniques optimize trajectories and identities for image batches (past and future). Computational effort in detection-free tracking depends on the number of attended targets at each time instant. In MOT with resource constraints scenarios, the observer's goal is to predict the best regions in the visual field to attend, in the quest to evaluate if they pertain to a given set of persons of interest, and thus to prune the visual search space by filtering out irrelevant image locations.

Classical object detection algorithms are based on exhaustive search, sliding window approaches, which operate over the full image space, in a sliding window manner and are typically inefficient and agnostic to top-down temporal context. When combined with fast bottom-up saliency-based approaches that generate object bounding box proposals, the overall detection process becomes more efficient [225], since regions that are unlikely to contain

objects are discarded for further processing. However, these approaches are agnostic to object dynamics, and are solely based on low-level visual features.

Resource-constrained adaptive sensing, is within a different line of research, and accounts for dynamical uncertain environments and noisy sensors for sequential decision making. The temporal integration of continuously gathered noisy detections is used to predict future environment states and decide, in a top-down manner, where to allocate the limited sensing resources, according to some task-related goal. It has been shown that adaptive sensing improves not only processing efficiency but also estimation robustness when compared to non-adaptive approaches [131]. Adaptive sensing problems can be formulated as POMDPs [4][33][37] that, depending on the way they compute the policies, belong to two different paradigms: Offline methods compute full policies before run time. Despite achieving remarkable performance in visual search tasks, these often require the evaluation of many possible situations, via backward induction, and hence take a considerable amount of time (e.g. hours). Online decision approaches avoid the computational burden of computing full policies for many situations, by departing from the current belief state and simulating future rewards for a finite planning horizon [177]. Within the online POMDP domain, the work closest to ours is the one in [37], which proposed a formulation for general adaptive sensing problems. The authors applied rollout techniques which are guaranteed to improve upon a provided base policy, that may be hard or impossible to compute. Rollout techniques evaluate the candidate actions, by running many Monte-Carlo simulations and returning the action with the best average outcome. In this work we rely on a different, widely known algorithm named MCTS [29], which has recently been given much attention by the Artificial Intelligence community due to its outstanding performance in the game Go [69]. MCTS combines tree search with randomized rolllout simulations, being ideal for decision making under uncertainty.

In this work, we propose a probabilistic framework which poses the multiple object tracking-by-detection problem as an on-line, resource-constrained decision making, aimed at minimizing the combined targets' state uncertainty, while coping with computational processing limitations (see Figure 5.1). Our framework relies on object detections with associated confidence measures, obtained from visual information, that are used to drive the observer's attentional focus during multiple object tracking. More specifically, we pose our decision framework within the POMDP domain in order to account for non-deterministic dynamics and partially observable states. The derived dynamic resource allocation decision process combines prior knowledge about the targets' state dynamics with accumulated probabilistic information provided from sequentially gathered observations, in order to optimize multiple target location estimation precision (i.e. minimize tracking uncertainty). In the proposed formulation, actions are taken from a low dimensional binary space. This allows for finding decision policies in real-time using on-line, tree-based, planning algorithms for finite horizon POMDPs [177].

The main contribution of this thesis in this topic is a novel framework for dynamic and resource-constrained target selection attended at each time instant, during MOT. The proposed approach is formulated as a resource-constrained POMDP and inspired by brain limitations in the number of targets and the total area of the visual field to be attended. More specifically, first we model the state-dependent uncertainty that arises during detection due to the discrete nature of the sliding window based detector. Then, we apply an online MCTS method to solve the planning problem in real-time. To our knowledge we are the first to apply an online tree-based POMDP solver in a probabilistic resource-constrained multiple object tracking scenario. The computational benefits of our methodology are demonstrated in a multiple person tracking scenario, by combining it with a state-of-the-art pedestrian detection algorithm [45]. However, we note that the proposed decision making pipeline can be combined with any general object detection algorithm.

5.3 Methodologies

A POMDP for general active sensing can be defined as a 6-element tuple $(\mathcal{X}, \mathcal{A}, \mathcal{Y}, T, O, R)$ where $\mathcal{X}$, $\mathcal{A}$ and $\mathcal{Y}$ denote the set of the possible environment states, perceptual actions and observations, respectively. State transitions are modeled as a Markov process and represented by the probability distribution function (pdf) $T(x_t, x_{t-1}) =$

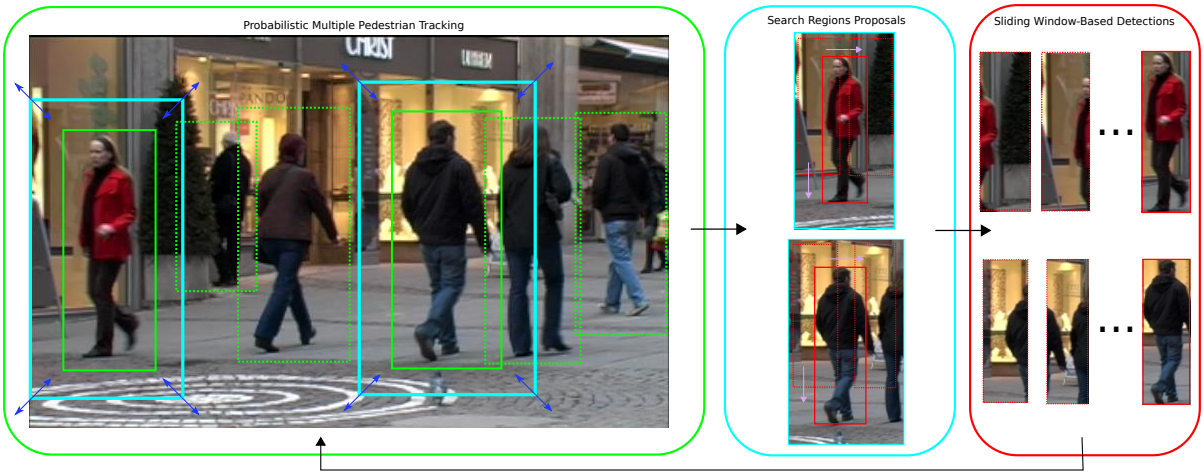


Figure 5.1: The proposed resource-constrained multiple pedestrian tracking pipeline. Given a set of persons being tracked, our decision making algorithm decides which sub-regions of the visual scene to attend. Then, a sliding window-based detector is applied to the selected search regions, instead of the whole image. For each region a winning candidate is obtained via maximum suppression and fed to the associated tracker with probabilistic measures queried from the observation model.

$p(x_t|x_{t-1})$. Observations are generated from states according to the pdf $O(x_t, a_t, y_t) = p(y_t|x_t, a_t)$. Under the resource-constrained adaptive sensing domain, the goal of the planning agent is to devise control strategies that generate perceptual actions from belief states, such that some intrinsic cumulative reward is maximized, while accounting for perceptual limitations. In the rest of this section we describe our resource-constrained POMDP formulation for multiple pedestrian tracking scenarios.

Let us consider a set of targets indexed by $\mathcal{K} = \{1, ..., K\}$, being tracked in a 2D image plane $\mathcal{I}$, with state $x^k \in \mathcal{X} \subset \mathbb{R}^3$ given by

$$x_t^k = \begin{bmatrix} x_t^{k,c} \\ x_t^{k,s} \end{bmatrix} \tag{5.1}$$

where $x^{k,c} = (u, v)$ and $x^{k,s}$ represent the bounding box centroid image coordinates and scale, respectively. Moreover, let us assume a stationary Markov chain $p(x_t^k|x_{t-1}^k)$ in order to model the object's state transition between consecutive frames. Similarly to [22] we assume sparsity-in-space and independence among targets, and linear constant-velocity dynamics model, which is a good approximation for targets that move with low acceleration in $3D$ and are not too close to the image plane. Finally, we assume that the targets' states are partially observable and statistically explained by the observation model distribution $p(y_t^k|x_t^k)$.

5.3.1 Recursive Bayesian Estimation

Object tracking can be achieved by means of recursive Bayesian estimation, according to

$$\begin{aligned} b_t^k &\stackrel{\text{def}}{=} p(x_t^k|y_{1:t}^k) \\ &= \eta p(y_t^k|x_t^k)\bar{b}_t^k \end{aligned} \tag{5.2}$$

where b_t^k represents the *belief* posterior probability over the target state x_t^k, given the set of all gathered observations $y_{1:t}^k$ taken up to time t, η is a normalizing factor and

$$\bar{b}_t^k = \int p(x_t^k|x_{t-1}^k) b_{t-1}^k dx_{t-1} \tag{5.3}$$

represents the belief after the prediction step. Furthermore, we assume Gaussian state transition and observation noises and hence tracking is optimally performed using K independent Kalman filters. At each time instant, each Kalman filter provides a parametric posterior probability distribution function (pdf) over the target state

$$b_t^k = \mathcal{N}(\hat{x}_t^k, \Sigma_t^k) \tag{5.4}$$

where

$$\hat{x}_t^k = \begin{bmatrix} \hat{x}_t^{k,c} \\ \hat{x}_t^{k,s} \end{bmatrix} \tag{5.5}$$

is the estimated state and

$$\Sigma_t^k = \begin{bmatrix} \sigma_t^{k,c} & 0 \\ 0 & \sigma_t^{k,s} \end{bmatrix} \tag{5.6}$$

is the error covariance matrix. Note that here we consider a diagonal covariance matrix and aggregate the centroid components in order to ease the notation.

5.3.2 Observation Model

The observations provided by the object detector are localized bounding boxes, obtained with a pedestrian detection algorithm. More specifically, at each time instant the agent collects a set of observations

$$\mathcal{Y}_t = \left\{ y_t^k, k = 1, ..., K \right\} \tag{5.7}$$

each corresponding to a noisy projection of the k target state.

Detection noise has several origins, the easiest to model being the one originated by the discrete nature of the detector. The noise affecting the center of a bounding box $\varepsilon_{x_t^{k,c}}$ has two origins, both depending on the scale of the bounding boxes: ε_{sl}, the error due to the sliding window process and ε_{sc_i}, the error due to the uncertainty of the size of the bounding box. The value of sliding window jumps Q_{sl} depends on the scale of the detection:

$$Q_{sl}^n = s^n Q_{sl}(0) \tag{5.8}$$

where Q_{sl}^n is the number of pixels between two consecutive sliding window positions at scale n and s^n is the value of scale n, defined as:

$$s^n = 2^{\frac{n}{N}} \tag{5.9}$$

where N is the number of scales per octave. The present implementation of the detector has $N = 8$.

The value of the jumps of the bounding box center-bottom due to scale change, depend on the scale. The number of pixels is given by Q_{scx} and Q_{scy}, for the x and y coordinates, respectively. In the worst case scenario, a jump from the actual scale to the coarsest one, these values are given by:

$$Q_{scx}^n = w^0(2^{\frac{n+1}{8}} - 2^{\frac{n}{8}})/2 \tag{5.10}$$

$$Q_{scy}^n = h^0(2^{\frac{n+1}{8}} - 2^{\frac{n}{8}})/2 \tag{5.11}$$

where w^0 and h^0 are the width and the height of the smallest bounding box ($n = 0$). Assuming a Gaussian distribution for these quantization errors, the statistics of ε_{sl} are given by

$$\mu_{sl}^n = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{\Sigma}_{sl}^n = \begin{bmatrix} (Q_{sl}^n)^2 & 0 \\ 0 & (Q_{sl}^n)^2 \end{bmatrix} \tag{5.12}$$

Regarding ε_{sc_i}, we approximate the statistics of these errors by the worst case which is given by

$$\mu_{sc}^n, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{\Sigma}_{sc}^n \approx \begin{bmatrix} (Q_{scx}^n)^2 & 0 \\ 0 & (Q_{scy}^n)^2 \end{bmatrix} \tag{5.13}$$

Since both sources of noise are independent but not additive, our observation model considers the largest one at each time. This yields the final image observation error ε^n:

$$\varepsilon^n \sim \mathcal{N}(\mathbf{0}, \Sigma^n) \tag{5.14}$$

where

$$\Sigma^n = \max(\Sigma_{sl}^n, \Sigma_{sc}^n) \tag{5.15}$$

5.3.3 Dynamic Search Regions

Let us now consider different time-varying (dynamic) regions of interest (i.e. bounding boxes) to be attended, each delimiting a target instance hypothesis

$$\mathbf{u}_t = \bigcup_{k \in \mathcal{K}} u_t^k \quad \text{where} \quad u_t^k \subset \mathcal{X} \tag{5.16}$$

Search regions are deterministically and analytically determined from beliefs according to the following mapping function

$$f : \hat{x}_t^k, \Sigma_t^k \rightarrow u_t^k \tag{5.17}$$

which is defined as follows

$$u_t^k = \left[\hat{x}_t^{k,c} - \alpha_c \sigma_t^{k,c}, \hat{x}_t^{k,c} + \alpha_c \sigma_t^{k,c}\right] \times \tag{5.18}$$

$$\left[\hat{x}_t^{k,s} - \alpha_s \sigma_t^{k,s}, \hat{x}_t^{k,s} + \alpha_s \sigma_t^{k,s}\right] \tag{5.19}$$

where α_s and α_c are user selected parameters that control the width of the confidence bounds and thus the size of the search regions. This definition accounts for the confidence level of the true target state being within the search region. The user selected parameters permit balancing the trade-off between accuracy and allocation effort (larger vs smaller regions).

Furthermore, we assume that each region has a deterministic, time-varying binary activation state

$$\mathcal{A} = \{a^k \in \mathbb{B}, k \in \mathcal{K}\} = \mathbb{B}^K \tag{5.20}$$

where $\mathbb{B} = \{0, 1\}$ with 0 and 1 meaning "not processing" and "processing", respectively. Decision making is therefore performed in a finite multi-dimensional binary action space and involves selecting which sub regions of the image space to apply the sliding window detector to perform measurement update steps. The belief becomes dependent on actions as follows

$$b_t^k(a_t^k) = \begin{cases} \bar{b}_t^k & \text{if } a_t^k = 0 \\ \eta p(y_t^k | x_t^k) \bar{b}_t^k & \text{if } a_t^k = 1 \end{cases} \tag{5.21}$$

where η is a normalizing constant. For attended regions, the predicted belief is approximated by the expected observation uncertainty given by the observation model, over a finite set of space points corresponding to detection

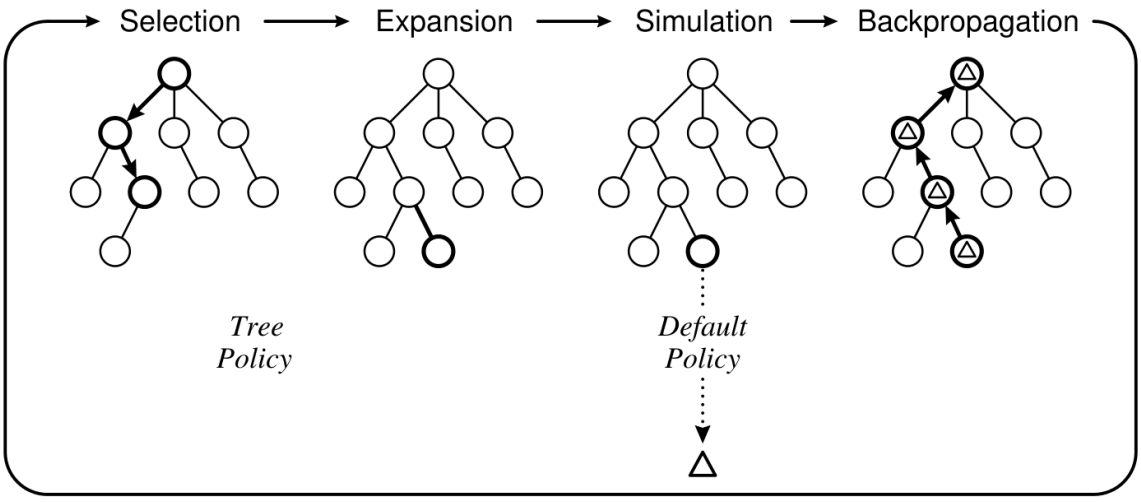


Figure 5.2: Monte Carlo Tree Search (image taken from [29]).

windows $\mathcal{Y}^k \subset \mathcal{X}$ in the search region k, according to

$$b_t^k(a_t^k) \approx c \sum_{i=1}^{|\mathcal{Y}^k|} p(y_t^k|x_t^{k,i})\bar{b}_t^{k,i} \quad \text{if } a_t^k = 1 \tag{5.22}$$

where c is a normalizing constant, $|\mathcal{Y}^k|$ is the number of detection windows and $\bar{b}_t^{k,i} = p(x_t^i|\bar{b}_t^k)$. Each $p(y_t^k|x_t^{k,i})$ is queried on-line from the learned observation model. Assessing multiple $x_t^i \in u_t^k$ instead of just $\hat{x}_t$ should better approximate the error distribution.

5.3.4 Resource constrained POMDP with belief-dependent rewards

As previously noted the decision making involved in resource constrained multiple target tracking scenarios can be formulated within the POMDP framework. The perceptual agent tries to minimize tracking uncertainty by prioritizing its limited attentional resources to promising image regions. The instantaneous reward function should thus reflect the action contribution to maximizing the information regarding the targets' states. Similarly to [7] let us define the instantaneous reward at time t as the negative entropy of the belief state, given by the following expectation

$$r(b_t^k(a_t^k)) = \int b_t^k \log b_t^k dx_t \tag{5.23}$$

For Gaussian beliefs this reward becomes simply given by

$$r(b_t^k(a_t^k)) \approx -\log(|\Sigma_t^k|) \tag{5.24}$$

Inspired by the evidence of visual processing capacity limitations in humans [220], we formulate the proposed resource constrained information maximization as follows:

$$
\begin{aligned}
\underset{a}{\text{maximize}} \quad & R_T = E\left[\sum_{\tau=1}^{T}\gamma^{\tau}\sum_{k=1}^{K} r(b_{t+\tau}^{k}(a_{t+\tau}^{k}))\right] \\
\text{subject to} \quad & \sum_{k=1}^{K} a_{t+\tau}^{k} \leq K_{\max} \quad \forall_{\tau\in\{1,\ldots,T\}} \\
& \sum_{k=1}^{K} a_{t+\tau}^{k} |u_{t+\tau}^{k}| \leq S_{\max} A_p \quad \forall_{\tau\in\{1,\ldots,T\}}
\end{aligned}
$$

where T is the planning horizon, $E[\cdot]$ is the expectation operation, $r(\cdot)$ is the reward function, $\gamma \in]0,1]$ is a discount factor, $|u_{t+\tau}^{k}|$ is the area of the k search region, $K_{\max}$ is the maximum region-based activation capacity, A_p is the image pixel area and $S_{\max}$ is the relative maximum image area that the visual system may process per time-instant. The first constraint reflects short-term memory limitations and allows reducing the action space (assuming $K_{\max} < K$), and thus the branching factor during planning. The second is motivated by computational effort and timing limitations that arise during visual processing and contributes to prune infeasible planning tree branches, by prioritizing resources to higher uncertainty targets.

5.3.5 Monte Carlo Tree Search (MCTS)

The MCTS algorithm relies on Monte-Carlo simulations to assess the nodes of a search tree in a best-first order, by prioritizing the expansion of the most promising nodes according to their expected reward. In a nutshell, the algorithm runs Monte Carlo simulations from the current belief state (i.e. input root node), and progressively builds a tree of belief states and outcomes. In the end, the most promising action is returned. Each run comprises four phases (see Figure 5.2):

1. Selection: In the selection step a sequence of actions are chosen within the search tree. Tree descending is performed from the root until a leaf node is reached. Action selection is typically carried out using an algorithm named UCB[112], which elegantly balances the exploration-exploitation trade-off, during action selection. On the one hand, based on the current accumulated simulated knowledge, the planning agent should select actions that may lead to the best immediate payoffs (exploitation). On the other hand, the agent should select unexplored actions since they may yield better long-term outcomes;

2. Expansion: an action that leads to an unvisited node is selected and the resulting expanded leaf node is appended to the tree;

3. Simulation: From the expanded node, actions are taken randomly in a Monte-Carlo depth-first manner, until a predefined horizon or a terminal state is reached. Simulation depth (i.e. time horizon) is typically fixed, to deal with real-time constraints. Since sampling from a uniform distribution over actions may be suboptimal, problem specific knowledge should be incorporated to give larger sampling probabilities to more promising actions. In our specific problem, we bias this sampling distribution such that regions with higher entropy are prioritized.

4. Back-propagation: Finally, the simulation rewards are back-propagated to the root node. This includes updating the reward rate stored at each node along the way.

Finally, runs are repeated until a computational budget (i.e. a triggering timeout or a maximum number of iterations) is reached, and the best action from the root node is selected.

Upper Confidence Bounds for Trees (UCT)

The idea of using Upper Confidence Bounds [10] on rewards to deal with the exploration exploitation dilemma in the face of uncertainty, has been widely applied to reinforcement learning problems. In MCTS, Upper Confidence

Bounds for Trees (UCT) are typically employed in the selection phase, while descending the tree. The upper confidence bound accounts for the currently estimated value of the action, and the estimated UCT variance, according to

$$UCT(a) = r + c\sqrt{\frac{\log n_v}{n_a}} \tag{5.25}$$

where r is the estimate for the value of the action based on the simulated payoffs, n_v is the number of times the node has been visited, and n_a is the number of times an action a has been tried from that node. The constant c is a problem-dependent parameter that balances the exploration-exploitation trade-off.

5.4 Experiments and Results

In order to evaluate the proposed resource-constrained tracking approach we performed a set of experiments on the TUD-Stadtmitte MOTChallenge dataset [121], which allows to evaluate tracking performance with the CLEAR MOT metrics and known ground truth [19]. This dataset comprises a video sequence of 179 images, acquired with a static camera with 640×480 image resolution. An average of 8 pedestrians are present in the visual field, during the video. To quantitatively assess the performance of our methodologies we focused our evaluation in the time speed-up gains and in the multiple object tracking precision (MOTP), which is the total error in estimated position for matched object-hypothesis pairs over all frames, averaged by the total number of matches:

$$\text{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \tag{5.26}$$

where $d_t^i \in [0, 100]$ quantifies the amount of overlap (in percentage) between the true object o_i and its associated hypothesis bounding boxes, and where c_t is the number of matches found for time t. The Multiple Object Tracking Precision (MOTP) shows the ability of the tracker to keep consistent trajectories.

Our aim was to investigate the performance of the proposed methodologies dependency on the resource-constraints. We considered the following activation capacities $K_{\text{max}} \in \{1, 2, 3, 4\}$ and maximum processing image areas $S_{\text{max}} \in [0.1, 1.0]$. Since the MCTS method is randomized, we performed 100 trials for each combination of parameters.

The region size parameters where found empirically and were set to $\alpha_c = \alpha_s = 1$. At each time step, the MCTS planning root node was set to the current tracking belief, and the algorithm was allowed to run for $10ms$. Finally, the simulation step depth was set to 3 and γ was set to 0.9. The association between detections and trackers was performed with the Hungarian Algorithm [30] using the Mahalanobis distance. The tracking process is bootstrapped in the first frame, by applying the pedestrian detector to the whole image and instantiating a tracker for each detection. These trackers are kept during the entire video sequence, and every non-assigned detection is discarded, i.e., trackers are not further created.

The results presented in figure 5.3 demonstrate that planning future resource allocations in a constrained setting, improves simultaneously detection times and tracking precision, when compared with the baseline, full-window detector. As illustrated by the temporal gain plots (first row of Figure 5.3), our method achieves detection times around 12 times faster than the baseline detector applied to the full-window (0.02 against 0.24 seconds, for $S_{\text{max}} = 0.1$), with comparable tracking performance. Furthermore, the MOTP metric results demonstrate that, on the one hand, constraining the attention to regions with high probability of pertaining a person, allows to improve detection accuracy and to reduce the possibility of erroneous detections in the targets' vicinities, which may lead to bad detection-tracker associations and hence degrade tracking precision. On the other hand, ignoring regions that are unlikely to contain a person allows to reduce the number of spurious wrong detections (i.e. false positives) that may also contribute to tracking performance degradation.

In conclusion, in the constrained setting the allocation of more computational resources yields better tracking precision, at the cost of increased computational effort. Therefore, depending on the application requirements, this trade-off can be easily balanced by carefully selecting the $K_{\max}$ and $S_{\max}$ resource-constraints.

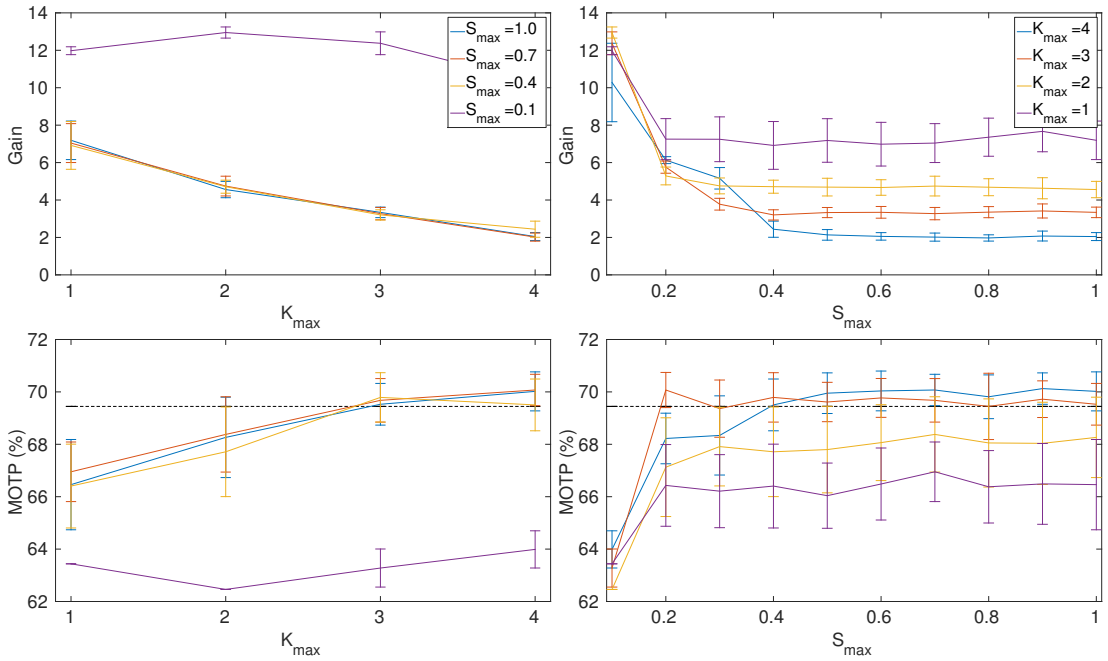


Figure 5.3: Speed-up gains and resulting Multiple Object Tracking Precision (MOTP). Bottom-row: Dashed black line represents the baseline full-window detector.

5.5 Conclusions

In this work we have addressed the MOT problem with constrained resources, which plays a fundamental role in the deployment of efficient mobile robots for real-time applications involved in HRI.

The tracking constraints are inspired by divided attention mechanisms in the brain. In particular, limitations in the number of targets and size of the attentional focci. We have framed the MOT within the POMDP domain in order to account for non-deterministic dynamics and partially observable states, and proposed a problem formulation that allows for on-line, real-time, planning with a state-of-the-art MCTS methodology.

Our framework relies on object detections and associated confidence measures, obtained from visual information, that are used to drive the observer's covert attentional foci during MOT. The derived process of decision making under uncertainty process combines prior knowledge about the targets' state dynamics with accumulated probabilistic evidence provided by sequentially gathered observations, in order to optimize multiple target location estimation precision (i.e. minimize tracking uncertainty). In the proposed formulation, actions are taken from a low dimensional binary space. This allows finding decision policies in real-time using on-line, tree-based, planning algorithms for finite horizon POMDPs. The results presented in this work show that directing the attentional focci to important image sub-regions allows for large detection speed-up improvements on tracking precision. Our contributions are the following:

1. to our knowledge, our work, published in publication , is the first to frame MOT as a resource-constrained POMDP problem, aimed at minimizing the combined targets' state uncertainty, while coping with computational processing limitations. The optimization constraints reflect experimental evidence on divided attention limits in the number and size of attended targets during MOT.

2. the results presented in this work demonstrate that constraining the attentional foci, in a top-down manner (in our case, a sliding window pedestrian detector) to image sub-regions and limiting the number of attended targets per time instant, allows for large tracking speed-ups and improvements on MOT precision.

Future work The major limitation of our approach is still its incapacity of dealing with non-sparse targets. In the future, data association should also be considered during planning by integrating data association methodologies such as Joint Probabilistic Data Association (JPDA) [77]. Another shortcoming of our methodology is its incapacity of locating new pedestrians appearing on the scene, in an efficient manner. However, this can be easily overcomed by considering proposals generated by bottom-up saliency methods. Finally, we note that the targets' dynamics and the observation distributions are extremely non-linear and non-Gaussian. Therefore, a mixture of particle filters [153] would be more appropriate for our particular problem, and hence improve tracking accuracy at the cost of some additional computational effort. Future research directions, may include:

- *Multi-object visual search, 3D pose estimation and tracking with biologically inspired binocular vision:* Extending the previously proposed target reconstruction pipeline with the ability of detecting and tracking the pose of multiple objects, using foveated stereo vision.

    1. *Resolving overlapping targets with foveated binocular disparity:* Target bottom-up detection overlaps could be resolved and tracklets initialized using real-time binocular disparity information.

    2. *Faster decision making:* in our current formulation, probabilistic state transition dynamics and observation models are known and assumed Gaussian, hence cheaper alternatives to MCTS look-ahead techniques, including greedy and beam search could be employed, and tracking performed in 3D.

    3. *Non-linear dynamic models for robust pose tracking:* the target dynamics and observation distributions are extremely non-linear and non-Gaussian when backprojected to 2D. Performing MOT in 3D using a mixture of particle filters [153] with particle proposals provided by our object detection and pose estimation approach, could be a more appropriate and robust choice for our particular problem to improve tracking accuracy, at the cost of some minor additional computational effort.

Chapter 6

Conclusions

In this thesis we have proposed approaches for biologically inspired artificial vision, ranging from low-level hardwired attention vision (i.e. foveal vision) to high-level visual attention mechanisms for robotics applications. More specifically, we delved beyond the state-of-the-art in biologically plausible space-variant resource-constrained vision methods (foveal vision, selective attention, active vision), for 2D recognition, 3D reconstruction, pose estimation and multiple object tracking tasks (see Table 6.1).

	2D recognition	3D reconstruction	Pose estimation	MOT
Foveal Vision	Chapter 2	Chapter 3		
Selective Attention	Chapter 2	Chapter 4	Chapter 4	
Active Vision	Chapter 2	Chapter 3		Chapter 5

Table 6.1: The biologically inspired vision mechanisms applied in this thesis to robotics relevant problems.

6.1 Foveal Vision

We have shown that **foveal vision** mechanisms are important both for monocular (2D) and binocular (3D) visual search tasks.

Monocular Vision We assessed the impact of mimicking non-uniform, human like, foveal vision mechanisms in recognition and localization tasks, when combined with state-of-the-art CNN architectures. In Chapter 2 we concluded that it is not necessary to store and transmit all the information present in high-resolution images. Although in the thesis we have not fully exploited the image compression effect of foveation, we have related the amount of information contained in foveal images to the size of the fovea. After a certain fovea size, roughly corresponding to half of the energy content of the original image, the performance in classification task saturates. The methodology is suitable for robotics agents that have the ability of moving the camera fixation point (i.e. pan and tilt), with low onboard computational resources, but with the ability to outsource processing, over low-bandwidth communication channels. When combined with saccadic mechanisms, our foveal representation can reach a recognition accuracy similar to the baseline high resolution image, with lower transmission bandwidth per saccade, than the original image.

Binocular Vision We demonstrated that foveal sensor topologies combined with stereo vision can be used to improve overall object reconstruction performance, at the cost of delayed task execution (i.e. target finding), when compared to conventional Cartesian counterparts (Chapter 3). We developed a novel way of characterizing the 3D stereo reconstruction error on foveal stereo setups.

6.2 Selective Attention

We have showed that biologically inspired selective attention mechanisms improve task execution and speed, in object detection and pose estimation problems. Namely, we addressed two types of selective attention mechanisms: shape-based attention and orientation selectivity.

Shape-Based Selective Attention We have proposed a biologically inspired pre-attentive architecture that filters out visual object distractors by smartly combining 3D saliency information with 2D appearance features extracted with neural networks optimized to filter task-irrelevant (e.g. ungraspable) object distractors (Chapter 4). Our experiments show that the proposed approach doubles the speed of object detection tasks in scenarios, with an average of 75% visible object distractors.

3D Orientation Selectivity We have developed a novel method to sample non-uniformly the orientation space. The application of this method to tasks such as 3D visual search (Chapter 3) and pose estimation (Chapter 4) has shown advantages with respect to the uniform resolution analogues in terms of coping with limited working-memory resources (i.e. limited number of memory cells) while improving performance. We have shown that biasing the sampling of the orientation space in accordance to prior task knowledge leads to more efficient, flexible and adaptable memory allocation and to more effective behaviours during task execution, namely a speed up of $30\times$ in a target reconstruction task (Chapter 3). We also showed that task-based orientation representations, implemented as randomized Hough orientation voting spaces, dramatically improve estimation robustness and accuracy on cylinder identification tasks, using 3D depth information. A systematic quantitative analysis of robustness to outliers and noise, validated our approach. In comparison to the state-of-the-art method of Rabbani [168], our method is able to cope with more than 150% data points not belonging to the object of interest (i.e. outliers), and more than 100% sensory noise levels ($\%$ with respect to object size) (Chapter 4).

6.3 Active Vision

We have applied active vision concepts to multiple problems, namely, for object recognition and 3D reconstruction, with foveal vision, and divided attention for MOT tasks.

Object Recognition In the case of 2D recognition, we developed neural saliency mechanisms to center objects within the fovea through saccades, and demonstrated similar recognition accuracy can be achieved with artificial foveal images when compared to full, high-resolution images (Chapter 2).

3D Object Recognition In 3D reconstruction, we have showed that by actively selecting the NBV point, using probabilistic evidence encoded in the proposed working-memory structure (i.e. SES), we were able to improve object reconstruction accuracy (Chapter 3). We showed that a MAB formulation using probabilistic 3D measures for NBV in 3D allows dealing effectively with exploration-exploitation trade-offs during 3D object reconstruction tasks, and also that different gaze patterns emerge depending on the sensor topology (i.e. Cartesian and Foveal) and field of views. The obtained results demonstrated that log-polar setups improve target reconstruction accuracy, at the cost of delayed task execution.

Multiple Object Tracking In MOT (Chapter 5), active vision was used to select the targets, and update their state over time, considering that at each frame there is a limited number of targets to track, and a limited budget for the targets and visual input to process. We have showned that we can reach similar tracking accuracy, while being 12 times faster than the baseline sliding-window based trackers.

6.4 Future Work

Although the proposed methodologies presented in this thesis are biologically plausible and exhibit many similarities with phenomena found in the human visual system, their main limitations, relies in the fact that they are purely model-based and do not apply learning principles. However, it is known that many processes in biological systems are driven by data and involve learning mechanisms that allow the agent to adapt to the environment.

Future work should target data-driven mechanisms that learn from brain and eye-tracking data to performe biomimetic saccadic ocular-motor control, in particular models that learn from human demonstrations how to explore the surrounding environment. In particular, the statistics of our reconfigurable short-term memory and orientation selectivity structures (Chapters 3 and 4) could be learned, using contextual data gathered from real-world scenarios (e.g. tabletop grasping scenarios).

While the proposed NBV planning methods allows to easily integrate task-dependent priors through biased sampling of the sphere of directions, we believe that its combination with conventional allocentric environment represention for localization and mapping, may be more appropriate for artificial agents that have the ability of moving their stereo apparatus freely in 6D space.

Finally, we also suggest extending and incorporating the developed divided attention mechanisms for MOT, within more sophisticated data-driven learning frameworks, such as [82], for real applications, such as surveillance and HRI.

Bibliography

[1] Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. "Pyramid methods in image processing". In: *RCA engineer* 29.6 (1984), pp. 33–41 (cit. on pp. 4, 6).

[2] Ankur Agarwal and Andrew Blake. "Dense stereo matching over the Panum band". In: *IEEE transactions on pattern analysis and machine intelligence* 32.3 (2010), pp. 416–430 (cit. on p. 50).

[3] Rajeev Agrawal. "Sample mean based index policies with O (log n) regret for the multi-armed bandit problem". In: *Advances in Applied Probability* (1995), pp. 1054–1078 (cit. on pp. 31, 43).

[4] Sheeraz Ahmad and Angela J. Yu. "Active Sensing as Bayes-Optimal Sequential Decision Making". In: *CoRR* abs/1305.6650 (2013). URL: `http://arxiv.org/abs/1305.6650` (cit. on pp. 32, 73).

[5] Emre Akbas and Miguel P Eckstein. "Object detection through search with a foveated visual system". In: *PLoS computational biology* 13.10 (2017), e1005743 (cit. on p. 4).

[6] Dima Amso and Gaia Scerif. "The attentive brain: insights from developmental cognitive neuroscience". In: *Nature Reviews Neuroscience* 16.10 (2015), pp. 606–619 (cit. on pp. 2, 56).

[7] Mauricio Araya, Olivier Buffet, Vincent Thomas, and Françcois Charpillet. "A POMDP extension with belief-dependent rewards". In: *Advances in Neural Information Processing Systems*. 2010, pp. 64–72 (cit. on p. 77).

[8] M Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking". In: *IEEE Transactions on signal processing* 50.2 (2002), pp. 174–188 (cit. on p. 50).

[9] Jean-Yves Audibert and Sébastien Bubeck. "Best arm identification in multi-armed bandits". In: *COLT-23th Conference on Learning Theory-2010*. 2010, 13–p (cit. on p. 42).

[10] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem". In: *Machine learning* 47.2-3 (2002), pp. 235–256 (cit. on pp. 30, 78).

[11] João Avelino, Rui Figureeiredo, Plinio Moreno, and Alexandre Bernardino. "On the perceptual advantages of visual suppression mechanisms for dynamic robot systems". In: *International Conference on Biologically Inspired Cognitive Architectures (BICA)*. 2016 (cit. on p. 45).

[12] Donald G Bailey and Christos-Savvas Bouganis. "Vision sensor with an active digital fovea". In: (2009), pp. 91–111 (cit. on p. 4).

[13] Sumitha L. Balasuriya. "A computational model of space-variant vision based on a self-organised artificial retina tessellation". PhD thesis. University of Glasgow, UK, 2006. URL: `http://theses.gla.ac.uk/4934/` (cit. on p. 4).

[14] Sumitha Balasuriya and Paul Siebert. "A biologically inspired computational vision front-end based on a self-organised pseudo-randomly tessellated artificial retina". In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* Vol. 5. IEEE. 2005, pp. 3069–3074 (cit. on p. 3).

[15] Dana H Ballard. “Generalizing the Hough transform to detect arbitrary shapes”. In: *Pattern recognition* 13.2 (1981), pp. 111–122 (cit. on p. 69).

[16] Miguel Jorge Bastos. “Modeling human gaze patterns to improve visual search in autonomous systems”. MA thesis. Instituto Superior Técnico, 2016 (cit. on p. 19).

[17] Momotaz Begum and Fakhri Karray. “Visual attention for robotic cognition: a survey”. In: *IEEE Transactions on Autonomous Mental Development* 3.1 (2011), pp. 92–105 (cit. on pp. 2, 4, 7).

[18] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. “Greedy Layer-Wise Training of Deep Networks”. In: *Advances in Neural Information Processing Systems* 19.1 (2007), p. 153. ISSN: 01628828. DOI: citeulike-article-id:4640046. arXiv: 0500581 [submit]. URL: http://papers.nips.cc/paper/3048-greedy-layer-wise-training-of-deep-networks.pdf (cit. on p. 101).

[19] Keni Bernardin and Rainer Stiefelhagen. “Evaluating multiple object tracking performance: the CLEAR MOT metrics”. In: *EURASIP Journal on Image and Video Processing* 2008.1 (2008), pp. 1–10 (cit. on p. 79).

[20] Alexandre Bernardino and José Santos-Victor. “A Binocular Stereo Algorithm for Log-Polar Foveated Systems”. English. In: *Biologically Motivated Computer Vision*. Ed. by HeinrichH. Bülthoff, Christian Wallraven, Seong-Whan Lee, and TomasoA. Poggio. Vol. 2525. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2002, pp. 127–136. ISBN: 978-3-540-00174-4 (cit. on p. 32).

[21] Alexandre Bernardino and José Santos-Victor. “Binocular tracking: integrating perception and control”. In: *IEEE Transactions on Robotics and Automation* 15.6 (1999), pp. 1080–1094 (cit. on p. 4).

[22] Alex Bewley, ZongYuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. “Simple Online and Realtime Tracking”. In: *CoRR* abs/1602.00763 (2016). URL: http://arxiv.org/abs/1602.00763 (cit. on p. 74).

[23] James W Bisley. “The neural basis of visual attention”. In: *The Journal of physiology* 589.1 (2011), pp. 49–57 (cit. on p. 8).

[24] Marc Bolduc and Martin D Levine. “A review of biologically motivated space-variant data reduction models for robotic vision”. In: *Computer vision and image understanding* 69.2 (1998), pp. 170–184 (cit. on p. 4).

[25] A. Borji and L. Itti. “State-of-the-Art in Visual Attention Modeling”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35.1 (Jan. 2013), pp. 185–207. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2012.89 (cit. on p. 7).

[26] A Borji and L Itti. “State-of-the-art in visual attention modelling”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 185–207 (cit. on pp. 2, 4, 11).

[27] Martim Brandao, Ricardo Ferreira, Kenji Hashimoto, José Santos-Victor, and Atsuo Takanishi. “Active Gaze Strategy for Reducing Map Uncertainty along a Path”. In: *3rd IFToMM International Symposium on Robotics and Mechatronics*. Oct. 2013, pp. 455–466 (cit. on p. 31).

[28] Donald Broadbent. *Perception and Communication*. Pergamon Press, 1958 (cit. on p. 7).

[29] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. “A survey of monte carlo tree search methods”. In: *IEEE Transactions on Computational Intelligence and AI in Games* 4.1 (2012), pp. 1–43 (cit. on pp. 73, 77).

[30] Rainer Burkard, Mauro Dell’Amico, and Silvano Martello. *Assignment Problems*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2009. ISBN: 0898716632, 9780898716634 (cit. on p. 79).

[31] Peter J. Burt and Edward H. Adelson. “The Laplacian Pyramid as a Compact Image Code”. In: *IEEE TRANSACTIONS ON COMMUNICATIONS* 31 (1983), pp. 532–540 (cit. on p. 19).

[32] Peter Burt and Edward Adelson. “The Laplacian pyramid as a compact image code”. In: *IEEE Transactions on communications* 31.4 (1983), pp. 532–540 (cit. on p. 6).

[33] Nicholas J Butko and Javier R Movellan. “Infomax control of eye movements”. In: *Autonomous Mental Development, IEEE Transactions on* 2.2 (2010), pp. 91–107 (cit. on pp. 32, 73).

[34] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Liang Wang, Chang Huang, Thomas S. Huang, Wei Xu, Deva Ramanan, and Yongzhen Huang. “Look and Think Twice : Capturing Top-Down Visual Attention with Feedback”. In: *IEEE International Conference on Computer Vision* (2015). DOI: 10.1109/ICCV.2015.338 (cit. on pp. 15, 17, 22, 105).

[35] Marisa Carrasco. “Visual attention: The past 25 years”. In: *Vision Research* 51.13 (2011), pp. 1484–1525. ISSN: 00426989. DOI: 10.1016/j.visres.2011.04.012 (cit. on p. 7).

[36] Patrick Cavanagh and George A Alvarez. “Tracking multiple targets with multifocal attention”. In: *Trends in cognitive sciences* 9.7 (2005), pp. 349–354 (cit. on p. 9).

[37] Edwin KP Chong, Christopher M Kreucher, and Alfred O Hero. “Monte-Carlo-based partially observable Markov decision process approximations for adaptive sensing”. In: *Discrete Event Systems, 2008. WODES 2008. 9th International Workshop on*. IEEE. 2008, pp. 173–180 (cit. on p. 73).

[38] Giorgia Committeri, Gaspare Galati, Anne-Lise Paradis, Luigi Pizzamiglio, Alain Berthoz, and Denis LeBihan. “Reference frames for spatial cognition: different brain areas are involved in viewer-, object-, and landmark-centered judgments about object location”. In: *Journal of Cognitive Neuroscience* 16.9 (2004), pp. 1517–1535 (cit. on p. 32).

[39] Nelson Cowan. “The magical number 4 in short-term memory: A reconsideration of mental storage capacity”. In: *Behavioral and Brain Sciences* 24.1 (2001), pp. 87–114. DOI: 10.1017/S0140525X01003922 (cit. on p. 9).

[40] Dennis D. Cox and Susan John. “SDO: A Statistical Method for Global Optimization”. In: *in Multidisciplinary Design Optimization: State-of-the-Art*. 1997, pp. 315–329 (cit. on p. 43).

[41] L Elizabeth Crawford, David Landy, and Amanda N Presson. “Bias in spatial memory: prototypes or relational categories”. In: *Poster presented at the 36th Annual Conference of the Cognitive Science Society, Quebec* (2014) (cit. on p. 34).

[42] Navneet Dalal and Bill Triggs. “Histograms of oriented gradients for human detection”. In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE. 2005, pp. 886–893 (cit. on p. 16).

[43] Atabak Dehban, Lorenzo Jamone, Adam R Kampff, and José Santos-Victor. “A deep probabilistic framework for heterogeneous self-supervised learning of affordances”. In: *IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids), 2017*. IEEE. 2017, pp. 476–483 (cit. on p. 55).

[44] Jeffrey Delmerico, Stefan Isler, Reza Sabzevari, and Davide Scaramuzza. “A comparison of volumetric information gain metrics for active 3D object reconstruction”. In: *Autonomous Robots* 42.2 (2018), pp. 197–208 (cit. on p. 31).

[45] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. “Fast Feature Pyramids for Object Detection”. In: *PAMI* (2014) (cit. on p. 73).

[46] Piotr Dollár, Ron Appel, and Wolf Kienzle. “Crosstalk Cascades for Frame-rate Pedestrian Detection”. In: *Proceedings of the 12th European Conference on Computer Vision - Volume Part II*. ECCV’12. Florence, Italy: Springer-Verlag, 2012, pp. 645–659. ISBN: 978-3-642-33708-6. DOI: 10.1007/978-3-642-33709-3_46. URL: http://dx.doi.org/10.1007/978-3-642-33709-3_46 (cit. on p. 16).

[47] Christian Dornhege and Alexander Kleiner. “A frontier-void-based approach for autonomous exploration in 3d”. In: *Advanced Robotics* 27.6 (2013), pp. 459–468 (cit. on p. 31).

[48] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. “Model globally, match locally: Efficient and robust 3D object recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee. 2010 (cit. on p. 57).

[49] Shimon Edelman. “Receptive fields for vision: From hyperacuity to object recognition”. In: (1995) (cit. on p. 30).

[50] Charles W Eriksen and James D St James. “Visual attention within and around the field of focal attention: A zoom lens model”. In: *Perception & psychophysics* 40.4 (1986), pp. 225–240 (cit. on p. 8).

[51] Mark Everingham, S. M Ali Eslami, Luc Van Gool, Christopher K I Williams, John Winn, and Andrew Zisserman. “The Pascal Visual Object Classes Challenge: A Retrospective”. In: *International Journal of Computer Vision* 111.1 (2014), pp. 98–136. ISSN: 15731405. DOI: 10.1007/s11263-014-0733-5 (cit. on p. 104).

[52] Karl Pearson F.R.S. “On lines and planes of closest fit to systems of points in space”. In: *Philosophical Magazine Series 6* 2.11 (1901), pp. 559–572. DOI: 10.1080/14786440109462720 (cit. on p. 62).

[53] J.F. Ferreira, Pierre Bessière, Kamel Mekhnacha, J. Lobo, J. Dias, and Christian Laugier. “Bayesian Models for Multimodal Perception of 3D Structure and Motion”. In: *International Conference on Cognitive Systems (CogSys 2008)*. Karlsruhe, Germany, 2008. URL: https://hal.archives-ouvertes.fr/hal-00338800 (cit. on p. 34).

[54] R. P. de Figureeiredo, Plinio Moreno, Alexandre Bernardino, and José Santos-Victor. “Multi-object detection and pose estimation in 3d point clouds: A fast grid-based bayesian filter”. In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2013, pp. 4250–4255 (cit. on p. 69).

[55] Rui Pimentel de Figureeiredo, Plinio Moreno, and Alexandre Bernardino. “Efficient pose estimation of rotationally symmetric objects”. In: *Neurocomputing* 150 (2015), pp. 126–135 (cit. on p. 57).

[56] Rui Figureeiredo, Plinio Moreno, and Alexandre Bernardino. “Robust cylinder detection and pose estimation using 3D point cloud information”. In: *IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. Apr. 2017 (cit. on pp. 56, 58).

[57] Rui Figureeiredo, Ashwini Shukla, Duarte Aragao, Plinio Moreno, Alexandre Bernardino, José Santos-Victor, and Aude Billard. “Reaching and grasping kitchenware objects”. In: *IEEE/SICE International Symposium on System Integration (SII)*. 2012, pp. 865–870 (cit. on p. 55).

[58] Martin A. Fischler and Robert C. Bolles. “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography”. In: *Commun. ACM* 24.6 (June 1981), pp. 381–395. ISSN: 0001-0782. DOI: 10.1145/358669.358692 (cit. on p. 57).

[59] Katherine Achim Fleming, Richard Alan Peters II, and Robert E. Bodenheimer. “Image Mapping and Visual Attention on a Sensory Ego-Sphere”. In: *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2006, October 9-15, 2006, Beijing, China*. 2006, pp. 241–246. DOI: 10.1109/IROS.2006.281688. URL: http://dx.doi.org/10.1109/IROS.2006.281688 (cit. on p. 33).

[60] Roland W Fleming. “Visual perception of materials and their properties”. In: *Vision research* 94 (2014), pp. 62–75 (cit. on p. 57).

[61] Steven L Franconeri, Jeffrey Y Lin, James T Enns, Zenon W Pylyshyn, and Brian Fisher. “Evidence against a speed limit in multiple-object tracking”. In: *Psychonomic bulletin & review* 15.4 (2008), pp. 802–808 (cit. on p. 9).

[62] Yoav Freund and Robert E Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”. In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139 (cit. on p. 16).

[63] Simone Frintrop. *"VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search"*. Springer-Verlag Berlin Heidelberg, 2006. ISBN: 978-3-540-32760-8 (cit. on p. 12).

[64] Karl Friston, Rick Adams, and Read Montague. “What is value – accumulated reward or evidence?” In: *Frontiers in Neurorobotics* 6.11 (2012). ISSN: 1662-5218. DOI: 10.3389/fnbot.2012.00011. URL: http://www.frontiersin.org/neurorobotics/10.3389/fnbot.2012.00011/abstract (cit. on p. 32).

[65] Viktor Gal, Lajos R Kozák, István Kóbor, Eva M Bankó, John T Serences, and Zoltán Vidnyánszky. “Learning to filter out visual distractors”. In: *European Journal of Neuroscience* 29.8 (2009), pp. 1723–1731 (cit. on p. 56).

[66] Dashan Gao and Nuno Vasconcelos. “Bottom-up saliency is a discriminant process”. In: *Proceedings of the IEEE International Conference on Computer Vision* (2007). ISSN: 1550-5499. DOI: 10.1109/ICCV.2007.4408851 (cit. on p. 11).

[67] Gabriel Garcı/a, Carlos Jara, Jorge Pomares, Aiman Alabdo, Lucas Poggi, and Fernando Torres. “A survey on FPGA-based sensor systems: towards intelligent and reconfigurable low-power sensors for computer vision, control and signal processing”. In: *Sensors* 14.4 (2014), pp. 6247–6278 (cit. on p. 4).

[68] Wilson S Geisler and Jeffrey S Perry. “Real-time foveated multiresolution system for low-bandwidth video communication”. In: *Photonics West'98 Electronic Imaging*. International Society for Optics and Photonics. 1998, pp. 294–305 (cit. on pp. 4, 7).

[69] Sylvain Gelly, Levente Kocsis, Marc Schoenauer, Michele Sebag, David Silver, Csaba Szepesvári, and Olivier Teytaud. “The grand challenge of computer Go: Monte Carlo tree search and extensions”. In: *Communications of the ACM* 55.3 (2012), pp. 106–113 (cit. on p. 73).

[70] R. Girshick. “Fast R-CNN”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015, pp. 1440–1448. DOI: 10.1109/ICCV.2015.169 (cit. on p. 17).

[71] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)* 9 (2010), pp. 249–256. ISSN: 15324435. DOI: 10.1.1.207.2059. URL: http://machinelearning.wustl.edu/mlpapers/paper%7B%5C_%7Dfiles/AISTATS2010%7B%5C_%7DGlorotB10.pdf (cit. on p. 104).

[72] Melvyn A Goodale and Angela Haffenden. “Frames of reference for perception and action in the human visual system”. In: *Neuroscience & Biobehavioral Reviews* 22.2 (1998), pp. 161–172 (cit. on p. 32).

[73] Lucian Cosmin Goron, Zoltan-Csaba Marton, Gheorghe Lazea, and Michael Beetz. “Robustly segmenting cylindrical and box-like objects in cluttered scenes using depth cameras”. In: *Robotics; Proceedings of ROBOTIK 2012; 7th German Conference on*. VDE. 2012, pp. 1–6 (cit. on p. 57).

[74] Stephen Gould, Joakim Arfvidsson, Adrian Kaehler, Benjamin Sapp, Marius Messner, Gary Bradski, Paul Baumstarck, Sukwon Chung, and Andrew Y. Ng. “Peripheral-foveal Vision for Real-time Object Recognition and Tracking in Video”. In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. IJCAI'07. Hyderabad, India: Morgan Kaufmann Publishers Inc., 2007, pp. 2115–2121. URL: http://dl.acm.org/citation.cfm?id=1625275.1625617 (cit. on p. 4).

[75] Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, and Jianwei Wan. “3D object recognition in cluttered scenes with local surface features: a survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.11 (2014), pp. 2270–2287 (cit. on p. 57).

[76] J-S Gutmann, Masaki Fukuchi, and Masahiro Fujita. “A floor and obstacle height map for 3D navigation of a humanoid robot”. In: *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*. IEEE. 2005, pp. 1066–1071 (cit. on p. 33).

[77] Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. “Joint probabilistic data association revisited”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 3047–3055 (cit. on p. 81).

[78] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. “Mask r-cnn”. In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE. 2017, pp. 2980–2988 (cit. on pp. 16, 56).

[79] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034 (cit. on p. 16).

[80] Christopher G Healey and James T Enns. “Attention and Visual Perception in Visualization and Computer Graphics”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.7 (2011), pp. 1–20 (cit. on p. 8).

[81] Geremy Heitz. *Graphical models for high-level computer vision*. Stanford University, 2009 (cit. on p. 70).

[82] David Held, Sebastian Thrun, and Silvio Savarese. “Learning to track at 100 fps with deep regression networks”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 749–765 (cit. on pp. 16, 56, 85).

[83] Martial Herbert, Claude Caillas, Eric Krotkov, In-So Kweon, and Takeo Kanade. “Terrain mapping for a roving planetary explorer”. In: *Robotics and Automation, 1989. Proceedings., 1989 IEEE International Conference on*. IEEE. 1989, pp. 997–1002 (cit. on p. 33).

[84] David A Hinkle and Charles E Connor. “Three-dimensional orientation tuning in macaque area V4”. In: *Nature neuroscience* 5.7 (2002), p. 665 (cit. on p. 9).

[85] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. “A Fast Learning Algorithm for Deep Belief Nets”. In: *Neural Computation* 18.7 (2006), pp. 1527–1554. ISSN: 0899-7667. DOI: 10.1162/neco.2006.18.7.1527. arXiv: 1111.6189v1. URL: http://www.ncbi.nlm.nih.gov/pubmed/16764513%7B%5C%%7D5Cnhttp://www.mitpressjournals.org/doi/abs/10.1162/neco.2006.18.7.1527 (cit. on p. 101).

[86] Makoto Hirose, Hidenori Furuhashi, Takeo Miyasaka, and Kazuo Araki. “Reconstruction of Range Data by Means of Geodesic Dome Type Data Structure”. In: *The Journal of the Institute of Image Electronics Engineers of Japan* 31.3 (2002), pp. 388–395. DOI: 10.11371/iieej.31.388 (cit. on p. 34).

[87] Heiko Hirschmuller. “Stereo processing by semiglobal matching and mutual information”. In: *IEEE Transactions on pattern analysis and machine intelligence* 30.2 (2008), pp. 328–341 (cit. on p. 45).

[88] Matthew D Hoffman, Eric Brochu, and Nando de Freitas. “Portfolio Allocation for Bayesian Optimization.” In: Citeseer (cit. on p. 43).

[89] Armin Hornung, Kai M. Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. “OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees”. In: *Autonomous Robots* (2013). Software available at http://octomap.github.com. DOI: 10.1007/s10514-012-9321-0. URL: http://octomap.github.com (cit. on p. 33).

[90] Armin Hornung, Kai M Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. “OctoMap: An efficient probabilistic 3D mapping framework based on octrees”. In: *Autonomous Robots* 34.3 (2013), pp. 189–206 (cit. on pp. 33, 50).

[91] Paul Hough. *Method and Means for Recognizing Complex Patterns*. U.S. Patent 3.069.654. Dec. 1962 (cit. on p. 57).

[92] Piers D Howe, Todd S Horowitz, Istvan Akos Morocz, Jeremy Wolfe, and Margaret S Livingstone. "Using fMRI to distinguish components of the multiple object tracking task". In: *Journal of Vision* 9.4 (2009), pp. 10–10 (cit. on p. 10).

[93] Deng Huang, Theodore T Allen, William I Notz, and Ning Zeng. "Global optimization of stochastic black-box systems via sequential kriging meta-models". In: *Journal of global optimization* 34.3 (2006), pp. 441–466 (cit. on p. 45).

[94] Jui-Ting Huang, Jinyu Li, and Yifan Gong. "An analysis of convolutional neural networks for speech recognition". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015), pp. 4989–4993 (cit. on p. 102).

[95] David H Hubel and Torsten N Wiesel. "Receptive fields and functional architecture of monkey striate cortex". In: *The Journal of physiology* 195.1 (1968), pp. 215–243 (cit. on p. 4).

[96] David H Hubel and Torsten N Wiesel. "Receptive fields of single neurones in the cat's striate cortex". In: *The Journal of physiology* 148.3 (1959), pp. 574–591 (cit. on p. 7).

[97] David H Hubel and Torsten N Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex". In: *The Journal of physiology* 160.1 (1962), pp. 106–154 (cit. on p. 7).

[98] Kai Huebner, Steffen Ruthotto, and Danica Kragic. "Minimum volume bounding box decomposition for shape approximation in robot grasping". In: *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE. 2008, pp. 1628–1633 (cit. on p. 57).

[99] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size". In: *arXiv:1602.07360* (2016) (cit. on pp. 60, 68, 69).

[100] Stefan Isler, Reza Sabzevari, Jeffrey Delmerico, and Davide Scaramuzza. "An information gain formulation for active volumetric 3d reconstruction". In: *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE. 2016, pp. 3477–3484 (cit. on p. 32).

[101] L. Itti and P. F. Baldi. "Bayesian Surprise Attracts Human Attention". In: *Advances in Neural Information Processing Systems, Vol. 19 (NIPS*2005)*. Cambridge, MA: MIT Press, 2006, pp. 547–554 (cit. on p. 32).

[102] Laurent Itti, Christof Koch, and Ernst Niebur. "A model of saliency-based visual attention for rapid scene analysis". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 11 (1998), pp. 1254–1259 (cit. on p. 11).

[103] Laurent Itti, Christof Koch, and Ernst Niebur. "A model of saliency-based visual attention for rapid scene analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.11 (1998), pp. 1254–1259. ISSN: 01628828. DOI: 10.1109/34.730558. arXiv: 0504378 [math] (cit. on pp. 12, 22).

[104] Maani Ghaffari Jadidi, Jaime Valls Miro, and Gamini Dissanayake. "Gaussian processes autonomous mapping and exploration for range-sensing mobile robots". In: *Autonomous Robots* 42.2 (2018), pp. 273–290 (cit. on p. 33).

[105] William James. "The principles of psychology (Vols. 1 & 2)". In: *New York Holt* 118 (1890), p. 688. DOI: 10.1037/10538-000 (cit. on p. 11).

[106] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. "Caffe: Convolutional architecture for fast feature embedding". In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 675–678 (cit. on pp. 24, 56, 103, 104).

[107] S.J. Julier and J.K. Uhlmann. "Unscented filtering and nonlinear estimation". In: *Proceedings of the IEEE* 92.3 (Mar. 2004), pp. 401–422. ISSN: 0018-9219. DOI: 10.1109/JPROC.2003.823141 (cit. on pp. 30, 38).

[108] T Kadir and J M Brady. "Scale, Saliency and Image Description". In: *International Journal of Computer Vision* 45.2 (2001), pp. 83–105. ISSN: 09205691. DOI: 10.1023/A:1012460413855 (cit. on p. 11).

[109] Fumi Katsuki and Christos Constantinidis. "Bottom-Up and top-down attention: different processes and overlapping neural systems". In: *The Neuroscientist* 20.5 (2014), pp. 509–521 (cit. on p. 2).

[110] D. J. Kerbyson and T. J. Atherton. "Circle detection using Hough transform filters". In: *Fifth International Conference on Image Processing and its Applications, 1995.* July 1995, pp. 370–374. DOI: 10.1049/cp:19950683 (cit. on p. 62).

[111] Christof Koch and Shimon Ullman. "Shifts in selective visual attention: towards the underlying neural circuitry". In: *Matters of intelligence*. Springer, 1987, pp. 115–141 (cit. on p. 11).

[112] Levente Kocsis and Csaba Szepesvári. "Bandit based monte-carlo planning". In: *European conference on machine learning*. Springer. 2006, pp. 282–293 (cit. on p. 78).

[113] Ioannis Kostavelis, Lazaros Nalpantidis, and Antonios Gasteratos. "Object recognition using saliency maps and htm learning". In: *Imaging Systems and Techniques (IST), 2012 IEEE International Conference on*. IEEE. 2012, pp. 528–532 (cit. on p. 58).

[114] Simon Kriegel, Christian Rink, Tim Bodenmüller, and Michael Suppa. "Efficient next-best-scan planning for autonomous 3D surface reconstruction of unknown objects". In: *Journal of Real-Time Image Processing* 10.4 (2015), pp. 611–631 (cit. on p. 32).

[115] David J Kriegman, Ernst Triendl, and Thomas O Binford. "Stereo vision and navigation in buildings for mobile robots". In: *IEEE Transactions on Robotics and Automation* 5.6 (1989), pp. 792–803 (cit. on pp. 30, 49).

[116] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012, pp. 1097–1105 (cit. on p. 16).

[117] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances In Neural Information Processing Systems* (2012), pp. 1–9 (cit. on pp. 24, 103–105).

[118] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105 (cit. on pp. 28, 69).

[119] Harold J Kushner. "A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise". In: *Journal of Fluids Engineering* 86.1 (1964), pp. 97–106 (cit. on p. 42).

[120] Tze Leung Lai and Herbert Robbins. "Asymptotically efficient adaptive allocation rules". In: *Advances in applied mathematics* 6.1 (1985), pp. 4–22 (cit. on p. 43).

[121] L. Leal-Taixe, A. Milan, I. Reid, S. Roth, and K. Schindler. "MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking". In: *arXiv:1504.01942 [cs]* (Apr. 2015). arXiv: 1504.01942. URL: http://arxiv.org/abs/1504.01942 (cit. on p. 79).

[122] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444. ISSN: 0028-0836. DOI: 10.1038/nature14539. arXiv: arXiv:1312.6184v5. URL: http://dx.doi.org/10.1038/nature14539 (cit. on pp. 100, 101).

[123] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context". In: *European conference on computer vision*. 2014, pp. 740–755 (cit. on p. 17).

[124] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. "Ssd: Single shot multibox detector". In: *European conference on computer vision*. Springer. 2016, pp. 21–37 (cit. on p. 16).

[125] Yong-Jin Liu, Jun-Bin Zhang, Ji-Chun Hou, Ji-Cheng Ren, and Wei-Qing Tang. "Cylinder detection in large-scale point cloud of pipeline plant". In: *IEEE transactions on visualization and computer graphics* 19.10 (2013), pp. 1700–1707 (cit. on pp. 57, 58).

[126] Daniel Lizotte, Tao Wang, Michael Bowling, and Dale Schuurmans. "Automatic gait optimization with gaussian process regression". In: *in Proc. of IJCAI*. 2007, pp. 944–949 (cit. on p. 43).

[127] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440 (cit. on p. 16).

[128] David G Lowe. "Object recognition from local scale-invariant features". In: *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Vol. 2. Ieee. 1999, pp. 1150–1157 (cit. on p. 16).

[129] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, Xiaowei Zhao, and Tae-Kyun Kim. "Multiple object tracking: A literature review". In: *arXiv preprint arXiv:1409.7618* (2014) (cit. on p. 72).

[130] Evelyne Lutton, Henri Maitre, and Jaime Lopez-Krahe. "Contribution to the determination of vanishing points using Hough transform". In: *IEEE transactions on pattern analysis and machine intelligence* 16.4 (1994), pp. 430–438 (cit. on pp. 60, 62).

[131] Matthew L Malloy and Robert D Nowak. "Near-optimal adaptive compressed sensing". In: *IEEE Transactions on Information Theory* 60.7 (2014), pp. 4001–4012 (cit. on p. 73).

[132] T. Mar, V. Tikhanoff, and L. Natale. "What can I do with this tool? Self-supervised learning of tool affordances from their 3D geometry." In: *IEEE Transactions on Cognitive and Developmental Systems* (2017). ISSN: 2379-8920. DOI: 10.1109/TCDS.2017.2717041 (cit. on p. 55).

[133] Donald Meagher. "Geometric modeling using octree encoding". In: *Computer graphics and image processing* 19.2 (1982), pp. 129–147 (cit. on p. 33).

[134] A Message, Andy Farke, Andy Farke, Andy Farke, Victoria M Arbour, Michael E Burns, Robert M Sullivan, Spencer G Lucas, Amanda K Cantrell, Thomas L Suazo, Jean-renaud Boisserie, Antoine Souron, Hassane Taïsso Mackaye, Andossa Likius, Patrick Vignaud, Michel Brunet, Melissa Tallman, Nina Amenta, Eric Delson, Stephen R Frost, Deboshmita Ghosh, and Zachary S Klukkert. "Artificial Inteligence". In: August (2014). ISSN: 0196-6553. DOI: 10.1002/ejoc.201200111. arXiv: arXiv:1011.1669v3 (cit. on pp. 101, 102, 104).

[135] Giorgio Metta, Giulio Sandini, David Vernon, Lorenzo Natale, and Francesco Nori. "The iCub humanoid robot: an open platform for research in embodied cognition". In: *Proceedings of the 8th workshop on performance metrics for intelligent systems*. ACM. 2008, pp. 50–56 (cit. on p. 4).

[136] Philipp Michel, Joel Chestnutt, James Kuffner, and Takeo Kanade. "Vision-guided humanoid footstep planning for dynamic environments". In: *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*. IEEE. 2005, pp. 13–18 (cit. on p. 33).

[137] L. Mihaylova, T. Lefebvre, H. Bruyninckx, K. Gadeyne, and J. De Schutter. "Active Sensing for Robotics - A Survey". In: *in Proc. 5 th Int'l Conf. On Numerical Methods and Applications*. 2002, pp. 316–324 (cit. on p. 72).

[138] Andrew T Miller, Steffen Knoop, Henrik I Christensen, and Peter K Allen. "Automatic grasp planning using shape primitives". In: *Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on*. Vol. 2. IEEE. 2003, pp. 1824–1829 (cit. on p. 55).

[139] G. A. Miller. "The magical number seven plus or minus two: some limits on our capacity for processing information." In: *Psychological review* 63 2 (1956), pp. 81–97 (cit. on p. 9).

[140] Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. "Object vision and spatial vision: two cortical pathways". In: *Trends in neurosciences* 6 (1983), pp. 414–417 (cit. on p. 8).

[141] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. "Recurrent models of visual attention". In: *Advances in neural information processing systems*. 2014, pp. 2204–2212 (cit. on pp. 16, 28).

[142] Jonas Mockus. "On Bayesian Methods for Seeking the Extremum". In: *Proceedings of the IFIP Technical Conference*. London, UK, UK: Springer-Verlag, 1974, pp. 400–404. ISBN: 3-540-07165-2. URL: http://dl.acm.org/citation.cfm?id=646296.687872 (cit. on p. 43).

[143] Camilla Mohlin, Kerstin Sandholm, Kristina N Ekdahl, and Bo Nilsson. "The link between morphology and complement in ocular disease". In: *Molecular immunology* 89 (2017), pp. 84–99 (cit. on p. 3).

[144] Plinio Moreno, Ricardo Nunes, Rui Figueeiredo, Ricardo Ferreira, Alexandre Bernardino, José Santos-Victor, Ricardo Beira, Luı/s Vargas, Duarte Aragão, and Miguel Aragão. "Vizzy: A humanoid on wheels for assistive robotics". In: *Robot 2015: Second Iberian Robotics Conference*. Springer International Publishing, 2016, pp. 17–28 (cit. on p. 44).

[145] Plinio Moreno, Ricardo Nunes, Rui Figueeiredo, Ricardo Ferreira, Alexandre Bernardino, José Santos-Victor, Ricardo Beira, Luı/s Vargas, Duarte Aragão, and Miguel Aragão. "Vizzy: A humanoid on wheels for assistive robotics". In: *Robot 2015: Second Iberian Robotics Conference*. Springer International Publishing, 2016, pp. 17–28 (cit. on p. 55).

[146] M. Muja and M. Ciocarlie. *Table Top Segmentation Package* (cit. on p. 59).

[147] Mervin E. Muller. "A Note on a Method for Generating Points Uniformly on N-dimensional Spheres". In: *Commun. ACM* 2.4 (Apr. 1959), pp. 19–20. ISSN: 0001-0782. DOI: 10.1145/377939.377946. URL: http://doi.acm.org/10.1145/377939.377946 (cit. on pp. 39, 62).

[148] Jiri Najemnik and Wilson S Geisler. "Optimal eye movement strategies in visual search". In: *Nature* 434.7031 (2005), pp. 387–391 (cit. on p. 32).

[149] Thuyen Ngo and BS Manjunath. "Saccade gaze prediction using a recurrent neural network". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017, pp. 3435–3439 (cit. on p. 28).

[150] A. Nurunnabi, Y. Sadahiro, and R. Lindenbergh. "ROBUST CYLINDER FITTING IN THREE-DIMENSIONAL POINT CLOUD DATA". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-1/W1 (2017), pp. 63–70. DOI: 10.5194/isprs-archives-XLII-1-W1-63-2017. URL: https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-1-W1/63/2017/ (cit. on pp. 57, 58).

[151] Simon T O'Callaghan and Fabio T Ramos. "Gaussian process occupancy maps". In: *The International Journal of Robotics Research* 31.1 (2012), pp. 42–62 (cit. on p. 33).

[152] Simon O'Callaghan, Fabio T Ramos, and Hugh Durrant-Whyte. "Contextual occupancy maps using Gaussian processes". In: *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE. 2009, pp. 1054–1060 (cit. on p. 33).

[153] Kenji Okuma, Ali Taleghani, Nando De Freitas, James J Little, and David G Lowe. "A boosted particle filter: Multitarget detection and tracking". In: *European Conference on Computer Vision*. Springer. 2004, pp. 28–39 (cit. on p. 81).

[154] W. Osberger and A.J. Maeder. “Automatic identification of perceptually important regions in an image”. In: *Proceeding of the Fourteenth International Conference on Pattern Recognition* 1 (1998), pp. 701–704. ISSN: 1051-4651. DOI: 10.1109/ICPR.1998.711240. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=711240 (cit. on p. 11).

[155] Piotr Ozimek, Nina Hristozova, Lorinc Balog, and Jan Paul Siebert. “A Space-Variant Visual Pathway Model for Data Efficient Deep Learning”. In: *Frontiers in Cellular Neuroscience* 13 (2019), p. 36 (cit. on pp. 27, 28).

[156] B Paláncz, JL Awange, A Somogyi, N Rehány, T Lovas, B Molnár, and Y Fukuda. “A robust cylindrical fitting to point cloud data”. In: *Australian Journal of Earth Sciences* 63.5 (2016), pp. 665–673 (cit. on p. 58).

[157] Daniela Pamplona and Alexandre Bernardino. “Smooth Foveal vision with Gaussian receptive fields”. In: *9th IEEE-RAS International Conference on Humanoid Robots, Humanoids 2009, Paris, France, December 7-10, 2009*. 2009, pp. 223–229. DOI: 10.1109/ICHR.2009.5379575. URL: http://dx.doi.org/10.1109/ICHR.2009.5379575 (cit. on pp. 30, 37, 43).

[158] Raja Parasuraman and Steven Yantis. *The attentive brain*. Mit Press Cambridge, MA, 1998 (cit. on pp. 2, 56).

[159] Ashok Kumar Patil, Pavitra Holi, Sang Keun Lee, and Young Ho Chai. “An adaptive approach for the reconstruction and modeling of as-built 3D pipelines from point clouds”. In: *Automation in Construction* 75 (2017), pp. 65–78 (cit. on p. 58).

[160] M. Perrollaz, A. Spalanzani, and D. Aubert. “Probabilistic representation of the uncertainty of stereo-vision and application to obstacle detection”. In: *Intelligent Vehicles Symposium (IV), 2010 IEEE*. June 2010, pp. 313–318. DOI: 10.1109/IVS.2010.5548010 (cit. on p. 36).

[161] Richard Alan Peters Ii, Kimberly A. Hambuchen, and Robert E. Bodenheimer. “The Sensory Ego-sphere: A Mediating Interface Between Sensors and Cognition”. In: *Auton. Robots* 26.1 (Jan. 2009), pp. 1–19. ISSN: 0929-5593. DOI: 10.1007/s10514-008-9098-3. URL: http://dx.doi.org/10.1007/s10514-008-9098-3 (cit. on pp. 33, 34).

[162] M.I. Posner. *Cognitive Neuroscience of Attention*. Guilford Press, 2012. ISBN: 9781609189853. URL: http://books.google.pt/books?id=8yjEjoS7EQsC (cit. on p. 3).

[163] MI I Posner. “Orienting of attention”. In: *Quarterly journal of experimental psychology* 32.1 (1980), pp. 3–25. ISSN: 0033-555X. DOI: 10.1080/00335558008248231. URL: http://www.tandfonline.com/doi/abs/10.1080/00335558008248231 (cit. on p. 11).

[164] Zenon W Pylyshyn and Ron W Storm. “Tracking multiple independent targets: Evidence for a parallel tracking mechanism”. In: *Spatial vision* 3.3 (1988), pp. 179–197 (cit. on p. 9).

[165] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 2017, pp. 77–85. DOI: 10.1109/CVPR.2017.16. URL: https://doi.org/10.1109/CVPR.2017.16 (cit. on p. 57).

[166] Morgan Quigley, Josh Faust, Tully Foote, and Jeremy Leibs. “ROS: an open-source Robot Operating System”. In: *ICRA workshop on open source software*. Vol. 3. 3.2. 2009 (cit. on p. 56).

[167] Philip Quinlan and Ben Dyson. “Attention: general introduction, basic models and data”. In: *Cognitive Psychology* (2008), pp. 271–311. ISSN: 00930415. DOI: 10.1136/ewjm.172.2.83 (cit. on p. 7).

[168] Tahir Rabbani and Frank Van Den Heuvel. “Efficient hough transform for automatic detection of cylinders in point clouds”. In: *ISPRS WG III/3, III/4* 3 (2005), pp. 60–65 (cit. on pp. 55–58, 60–64, 66, 69, 84).

[169] Marc aurelio Ranzato, Christopher Poultney, Sumit Chopra, Yann L. Cun, Marc'aurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann L. Cun. "Efficient Learning of Sparse Representations with an Energy-Based Model". In: *Advances in Neural Information Processing Systems* 19.1 (2007), pp. 1137–1144. ISSN: 10495258. URL: http://papers.nips.cc/paper/3112-efficient-learning-of-sparse-representations-with-an-energy-based-model%7B%5C%%7D5Cnhttp://papers.nips.cc/paper/3112-efficient-learning-of-sparse-representations-with-an-energy-based-model.pdf (cit. on p. 101).

[170] Babak Rasolzadeh, Alireza Tavakoli Targhi, and Jan-Olof Eklundh. "An attentional system combining top-down and bottom-up influences". In: *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint Lecture Notes in Computer Science* 4840 (2007), pp. 123–140. ISSN: 03029743. DOI: 10.1007/978-3-540-77343-6_8. URL: http://www.springerlink.com/index/682P7080741754X3.pdf (cit. on p. 12).

[171] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788 (cit. on p. 16).

[172] S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (June 2017), pp. 1137–1149. ISSN: 0162-8828. DOI: 10/gc7rmb (cit. on p. 17).

[173] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., 2015, pp. 91–99. URL: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf (cit. on p. 68).

[174] Herbert Robbins et al. "Some aspects of the sequential design of experiments". In: *Bulletin of the American Mathematical Society* 58.5 (1952), pp. 527–535 (cit. on p. 42).

[175] Alessandro Roncone, Ugo Pattacini, Giorgio Metta, and Lorenzo Natale. "A Cartesian 6-DoF Gaze Controller for Humanoid Robots." In: *Robotics: science and systems*. Vol. 2016. 2016 (cit. on p. 4).

[176] Ari Rosenberg, Noah J Cowan, and Dora E Angelaki. "The visual representation of 3D object orientation in parietal cortex". In: *Journal of Neuroscience* 33.49 (2013), pp. 19352–19361 (cit. on p. 9).

[177] Stéphane Ross, Joelle Pineau, Sébastien Paquet, and Brahim Chaib-Draa. "Online planning algorithms for POMDPs". In: *Journal of Artificial Intelligence Research* 32 (2008), pp. 663–704 (cit. on p. 73).

[178] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer. "Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub". In: *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. May 2008, pp. 962–967. DOI: 10.1109/ROBOT.2008.4543329 (cit. on p. 33).

[179] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y (cit. on pp. 25, 60, 103).

[180] Radu Bogdan Rusu, Nico Blodow, Zoltan Csaba Marton, and Michael Beetz. "Close-range scene segmentation and reconstruction of 3D point cloud maps for mobile manipulation in domestic environments". In: *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2009, pp. 1–6 (cit. on p. 59).

[181] José Santos-Victor and Alexandre Bernardino. “Vision-based navigation, environmental representations and imaging geometries”. In: *Robotics Research*. Springer, 2003, pp. 347–360 (cit. on p. 4).

[182] Ruwen Schnabel, Roland Wahl, and Reinhard Klein. “Efficient RANSAC for point-cloud shape detection”. In: *Computer graphics forum*. Vol. 26. 2. Wiley Online Library. 2007, pp. 214–226 (cit. on p. 57).

[183] Eric L Schwartz. “Spatial mapping in the primate sensory projection: analytic structure and relevance to perception”. In: *Biological cybernetics* 25.4 (1977), pp. 181–194 (cit. on p. 4).

[184] Eric L Schwartz, Douglas N Greve, and Giorgio Bonmassar. “Space-variant active vision: definition, overview and examples”. In: *Neural Networks* 8.7-8 (1995), pp. 1297–1308 (cit. on p. 4).

[185] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. “Taking the human out of the loop: A review of bayesian optimization”. In: *Proceedings of the IEEE* 104.1 (2016), pp. 148–175 (cit. on p. 33).

[186] C. E. Shannon. “Communication In The Presence Of Noise”. In: *Proceedings of the IEEE* 86 (1998), pp. 447–457 (cit. on p. 21).

[187] Christian Siagian and Laurent Itti. “Rapid biologically-inspired scene classification using features shared with visual attention”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.2 (2007), pp. 300–312. ISSN: 01628828. DOI: 10.1109/TPAMI.2007.40 (cit. on p. 11).

[188] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: *Computer Vision and Pattern Recognition* (2014). arXiv: 1312.6034. URL: http://arxiv.org/abs/1312.6034 (cit. on pp. 18, 22, 105).

[189] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *International Conference on Learning Representations* (2015), pp. 1–14. ISSN: 09505849. DOI: 10.1016/j.infsof.2008.09.005. arXiv: 1409.1556. URL: http://arxiv.org/abs/1409.1556 (cit. on pp. 25, 104, 105).

[190] Marie Sjölinder. “Spatial cognition and environmental descriptions”. In: *Exploring navigation: towards a framework for design and evaluation of navigation in electronic spaces* (1998), pp. 47–58 (cit. on p. 32).

[191] E.N. Sokolov and O.S. Vinogradova. *Neuronal mechanisms of the orienting reflex*. L. Erlbaum Associates, 1975. ISBN: 9780470925621. URL: https://books.google.pt/books?id=T1Z9AAAAIAAJ (cit. on p. 11).

[192] Eric Sommerlade and Ian Reid. “Probabilistic surveillance with multiple active cameras”. In: *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE. 2010, pp. 440–445 (cit. on p. 72).

[193] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. “Dropout : A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research (JMLR)* 15 (2014), pp. 1929–1958. ISSN: 15337928. DOI: 10.1214/12-AOS1000. arXiv: 1102.4807 (cit. on p. 101).

[194] Bobby Boge Stojanoski and Matthias Niemeier. “Late electrophysiological modulations of feature-based attention to object shapes”. In: *Psychophysiology* 51.3 (2014), pp. 298–308 (cit. on p. 56).

[195] Yun-Ting Su and James Bethel. “Detection and robust estimation of cylinder features in point clouds”. In: *ASPRS Conference*. 2010 (cit. on p. 58).

[196] Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. 1st. Cambridge, MA, USA: MIT Press, 1998. ISBN: 0262193981 (cit. on p. 41).

[197] C Szegedy, W Liu, Y Jia, and P Sermanet. “Going deeper with convolutions”. In: *Computer Vision Foundation* (2014). DOI: 10.1109/CVPR.2015.7298594 (cit. on pp. 24, 104, 105).

[198] Saime Tek, Gul Jaffery, Lauren Swensen, Deborah Fein, and Letitia R Naigles. "The shape bias is affected by differing similarity among objects". In: *Cognitive development* 27.1 (2012), pp. 28–38 (cit. on p. 57).

[199] Steven P Tipper, Jon Driver, and Bruce Weaver. "Object-centred inhibition of return of visual attention". In: *The Quarterly Journal of Experimental Psychology* 43.2 (1991), pp. 289–298 (cit. on p. 11).

[200] Beau Tippetts, Dah Jye Lee, Kirt Lillywhite, and James Archibald. "Review of stereo vision algorithms and their suitability for resource-limited systems". In: *Journal of Real-Time Image Processing* 11.1 (2016), pp. 5–25 (cit. on p. 36).

[201] Trung-Thien Tran, Van-Toan Cao, and Denis Laurendeau. "Extraction of cylinders and estimation of their parameters from point clouds". In: *Computers & Graphics* 46 (2015), pp. 345–357 (cit. on pp. 57, 58).

[202] V Javier Traver and Alexandre Bernardino. "A review of log-polar imaging for visual perception in robotics". In: *Robotics and Autonomous Systems* 58.4 (2010), pp. 378–398 (cit. on p. 4).

[203] Anne Treisman. "Preattentive processing in vision". In: *Computer Vision, Graphics, and Image Processing* 31.2 (Aug. 1985), pp. 156–177. ISSN: 0734189X. DOI: 10.1016/S0734-189X(85)80004-9. URL: http://www.sciencedirect.com/science/article/pii/S0734189X85800049 (cit. on p. 8).

[204] Anne M Treisman. "A Feature-Integration Theory of Attention". In: *Cognitive Psychology* 12 (1980), pp. 97–136. ISSN: 00100285. DOI: 10.1016/0010-0285(80)90005-5 (cit. on p. 8).

[205] Rudolph Triebel, Patrick Pfaff, and Wolfram Burgard. "Multi-level surface maps for outdoor terrain mapping and loop closing". In: *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. IEEE. 2006, pp. 2276–2282 (cit. on p. 33).

[206] John K Tsotsos. "Analyzing vision at the complexity level". In: *Behavioral and brain sciences* 13.3 (1990), pp. 423–445 (cit. on p. 3).

[207] John K. Tsotsos. "The Complexity of Perceptual Search Tasks". In: *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'89. Detroit, Michigan: Morgan Kaufmann Publishers Inc., 1989, pp. 1571–1577. URL: http://dl.acm.org/citation.cfm?id=1623891.1624005 (cit. on p. 31).

[208] Ken-Ichiro Tsutsui, Masato Taira, and Hideo Sakata. "Neural mechanisms of three-dimensional vision". In: *Neuroscience Research* 51.3 (2005), pp. 221–229 (cit. on p. 8).

[209] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. "Selective Search for Object Recognition". In: *International Journal of Computer Vision* 104.2 (2013), pp. 154–171. ISSN: 1573-1405. DOI: 10/gddfjz (cit. on p. 17).

[210] Sethu Vijayakumar, Jörg Conradt, Tomohiro Shibata, and Stefan Schaal. "Overt visual attention for a humanoid robot". In: *Intelligent Robots and Systems, 2001. Proceedings. 2001 IEEE/RSJ International Conference on*. Vol. 4. IEEE. 2001, pp. 2332–2337 (cit. on p. 4).

[211] Hermann Von Helmholtz. *Handbuch der physiologischen Optik*. Vol. 9. 1866 (cit. on p. 7).

[212] Jenny R Wang and Nandan Parameswaran. "Survey of sports video analysis: research issues and applications". In: *Proceedings of the Pan-Sydney area workshop on Visual information processing*. Australian Computer Society, Inc. 2004, pp. 87–90 (cit. on p. 72).

[213] Jianhua Wang and Yuncai Liu. "A Closed-Form Solution of Reconstruction from Nonparallel Stereo Geometry Used in Image Guided System for Surgery". English. In: *Multimedia Content Analysis and Mining*. Ed. by Nicu Sebe, Yuncai Liu, Yueting Zhuang, and ThomasS. Huang. Vol. 4577. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, pp. 371–380 (cit. on p. 38).

[214] Z Wang. *Rate scalable foveated image and video communications [Ph. D. thesis]*. 2003 (cit. on p. 4).

[215] Martin Weier, Michael Stengel, Thorsten Roth, Piotr Didyk, Elmar Eisemann, Martin Eisemann, Steve Grogorick, André Hinkenjann, Ernst Kruijff, Marcus Magnor, et al. "Perception-driven Accelerated Rendering". In: *Computer Graphics Forum*. Vol. 36. 2. Wiley Online Library. 2017, pp. 611–643 (cit. on p. 3).

[216] Carl F. R. Weiman. "Binocular stereo via log-polar retinas". In: ed. by SPIE. 1995 (cit. on p. 44).

[217] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. "A survey of transfer learning". In: *Journal of Big Data* 3.1 (2016), p. 9. ISSN: 2196-1115. DOI: 10.1186/s40537-016-0043-6 (cit. on p. 60).

[218] J M Wolfe, K R Cave, and S L Franzel. "Guided search: an alternative to the feature integration model for visual search". In: *Journal of Experimental Psychology: Human Perception and Performance* 15.3 (1989), pp. 419–433. ISSN: 0096-1523. DOI: 2527952 (cit. on p. 11).

[219] Jeremy M Wolfe. "Guided Search 2 . 0 A revised model of visual search". In: *Psychnomic Bulletin & Review* 1.2 (1994), pp. 202–238. ISSN: 1531-5320. DOI: 10.3758/BF03200774 (cit. on p. 11).

[220] Yaoda Xu and Marvin M Chun. "Selecting and perceiving multiple visual objects". In: *Trends in cognitive sciences* 13.4 (2009), pp. 167–174 (cit. on p. 77).

[221] B. Yamauchi. "A frontier-based approach for autonomous exploration". In: *Computational Intelligence in Robotics and Automation, 1997. CIRA'97., Proceedings., 1997 IEEE International Symposium on*. July 1997, pp. 146–151. DOI: 10.1109/CIRA.1997.613851 (cit. on p. 31).

[222] Steven Yantis and E Jones. "Mechanisms of attentional selection: temporally modulated priority tags". In: *Perception and psychophysics* 50 (Sept. 1991), pp. 166–78 (cit. on p. 9).

[223] Sergey Zagoruyko and Nikos Komodakis. "Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer". In: *International Conference on Learning Representations*. 2017 (cit. on p. 56).

[224] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. "SUN: A Bayesian framework for saliency using natural statistics". In: *Journal of vision* 8.7 (2008), pp. 32–32 (cit. on p. 12).

[225] C. Lawrence Zitnick and Piotr Dollár. "Edge Boxes: Locating Object Proposals from Edges". In: *ECCV*. 2014 (cit. on p. 72).

Appendix A

Artificial Neural Networks

ANN are computational models inspired by the central nervous system of mammals, that try to mimic the way the brain solves problems. A neural network is an approximator that receives stimuli inputs, and maps them to an output. Its key element is its ability to learn implicit mapping functions between inputs and outputs, making it capable of recognizing complex patterns, and a powerful machine learning tool. Neural networks are organized in layers that establish connections between neurons. Each connection between two neurons has a weight that controls how the activation of the first neuron influences the second. at controls how the activation of the first neuron influences the second. The input units receive information from the outside world and communicate with one or more hidden layers where actual processing takes place. In classification networks, the hidden layers apply a distortion of the input data in a non-linear way with the aim of having linearly separable categories at the end [122]. The last hidden layer links the output layer where items are assigned to the believed belonging class (see Figure A.1). All neurons in the hidden layers are processed by an activation function that can be a linear, threshold or sigmoid function.

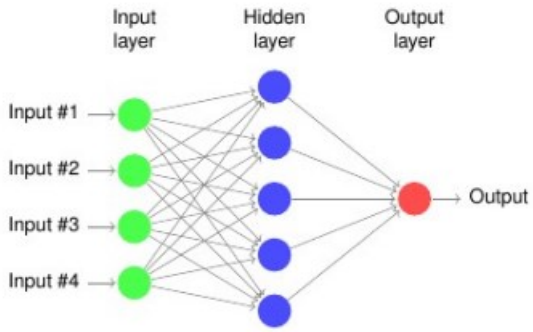


Figure A.1: Neural network basic structure.

There are two main learning algorithms for training neural network based classifiers:

- Supervised learning - requires a large labeled data set with labeled input samples. The network produces an output in the form of a vector of scores, one score for each category. Then, an objective function is computed to measure the error, i.e. the difference between the output scores and the desired pattern of scores. With this knowledge, all internal weights are adjusted with the goal of minimizing the error. To correctly perform these adjustments, the learning algorithm computes a gradient vector that, for each weight, indicates what would be the error value variation if the weight was increased by a small amount [122]. Finally, the weight vector is adjusted in the opposite direction to the gradient vector.
- Unsupervised learning - The network learns intrinsic relations about the data without specifying a target or label. It exploits only the statistical distribution of the input data to associate samples to groups of related

elements.

In supervised learning, there are mainly three steps to follow: the training set used to build the model by finding relationships between data and pre-classified targets (labeled data), the validation set is used to tune the hyper parameters such as the number of hidden units or the depth of the neural network and finally, the test set is used to estimate the performance of the model on never-seen data.

Deep Neural Networks

DNN are a subclass of ANN that are characterized by having several hidden layers between the input and output layers. Before 2006, most neural networks typically used one hidden layer, two at the most, due to the expensive cost of computation and the scarce amount of available data. The deep breakthrough occurred exactly in that year, 2006 when Hinton [85], Bengio [18] and Ranzato [169], three researchers brought together by the Canadian Institute for Advanced Research (CIFAR) were capable of training networks with much more layers for the handwriting recognition task. They used unsupervised learning methods to create layers of feature detectors without the need of labelled data. The deep breakthrough occurred in 2006 when Hinton [85], Bengio [18] and Ranzato [169], researchers brought together by the Canadian Institute for Advanced Research (CIFAR) were capable of training networks with much more layers for the handwriting recognition task [122].

Then, they pre-trained some layers with more complex feature detectors providing enough information to initialize the network weights with reasonable values. This method allowed researchers to train networks 10 or 20 times faster [122]. In recent years, DNNs are becoming deeper which resulted in a performance boost. However, very wide and shallow networks exhibit very weak performance at generalization despite being good at memorization. As opposed, deeper networks can learn features at several levels of abstraction and present much better results in generalization because they learn all the intermediate features between the raw data and the high-level classification. Note that using wider and deeper networks lead to an increase in the number of the parameters that the network will have to learn. Following the tendency to work with deeper networks and considering the overfitting problem that occurs when the model fits too closely to the data set, a recent technique called Dropout has been successfully implemented. The dropout technique consists on randomly dropping out (i.e. ignoring) neurons during the training phase [193] which enforces the network to learn more robust features and decrease co-dependency between neurons, improving the generalization of the neural network. One attempt to speed up the network by decreasing the number of parameters has been done by substituting large convolutions with the combination of smaller ones. Researchers replaced a large 7×7 convolution by a cascade of several small convolutions like 3×3 convolutions with the same depth [134]. In-between each of these small convolution layers, a ReLU layer is placed to increase the number of non-linearities. Therefore, we end up with a similar network but with fewer weights that result in fewer computations and a faster network. However, this type of substitution can not be done on the first layer because it will result in an enormous consumption of memory [134].

Convolutional Neural Networks

There are several types of DNNs but as far as visual classification and object detection is concerned, the most commonly used are the CNNs, that are *feed-forward* ANNs that take into account the spatial structure of the input. They have the ability to learn discriminative features from raw data input and have been used in several visual tasks including object recognition and classification. This type of neural networks is named convolutional since they perform the mathematical operation *convolution*. The mathematical formula for convolution of discrete signals is defined in (A.1) where x is the input signal and h is the impulse response. This operation has several applications on signal processing such as filtering signals (2D - image processing) or finding patterns between them.

$$y[n] = x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k]. \tag{A.1}$$

A CNN is constituted by multiple stacked layers that filter (convolve) the input stimuli to extract useful and meaningful information depending on the task at hand. These layers have parameters that are learned in a way that allows filters to automatically adjust to extract useful information without feature selection so there is no need to manually select relevant features. The general architecture of a CNN is shown in Figure A.2.

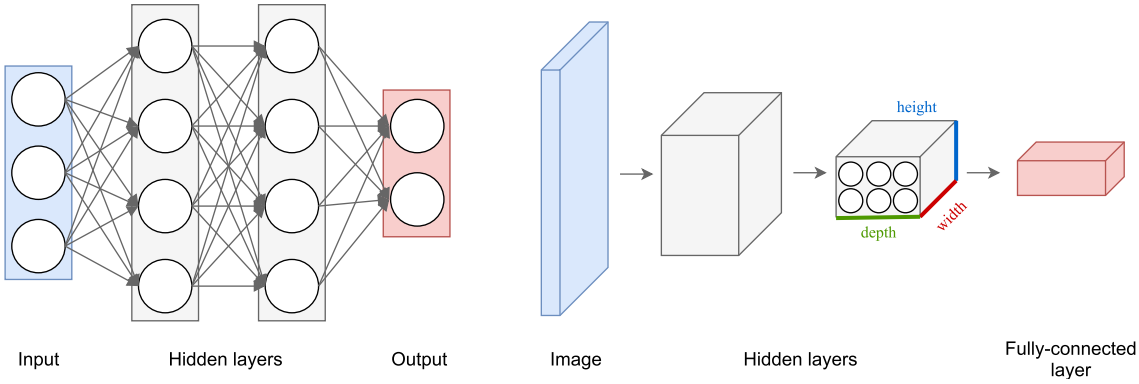


Figure A.2: Convolutional Neural Network architecture. Figure adapted from [94].

Convolutional layer: Each neuron receives a sub-region from a previous layer as input and these local inputs are multiplied by the weights. These filters are applied throughout the input space with the purpose of looking for specific features. Their weights are shared and their output is a feature map.

To configure a convolution layer, it is necessary to set some hyper parameters [134] such as:

- Kernel size - size of the filters;
- Stride - number of pixels that the kernel window will slide (usually, 1 for convolution layers);
- Number of filters - number of patterns that the convolution layer will look for.

Pooling layer: Is generally placed in-between convolutional layers and their goal is to downsample the input, reduce the dimensionality and produce a single output from the local region. It also decreases the amount of computation in the upstream layers by reducing the number of parameters to learn and provides basic translation invariance. A commonly used down-sampling function is the max-pooling which determines the maximum value within each sub-region (see Figure A.3.)

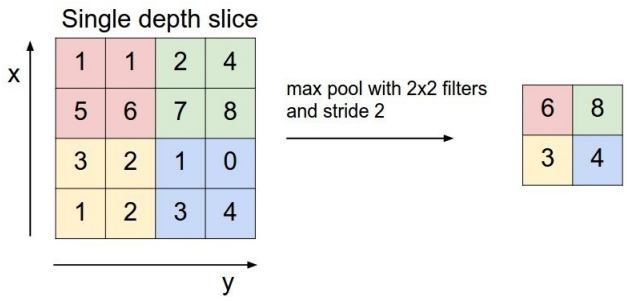


Figure A.3: Representation of max-pooling operation.[1]

[1]source: https://www.quora.com/What-is-max-pooling-in-convolutional-neural-networks [seen in December, 2016]

Fully-connected layer: Is the upper layer and computes the class scores to be consistent with training set labels. The input of the fully-connected layers corresponds to the set of all feature maps from the previous layer. Since they are not spatially located, there may be only be a convolutional layer after a fully-connected one.

In a CNN, the neurons are arranged in a 2D structure (width, height) in a way that allows spatial relations between neurons and original data to be preserved. However, with the use of colored images specially RGB images, an additional dimension for separate color channels is required. In this way, we have a 3D input (width, height and depth). The number of input neurons residing in the first network layer is equal to the input size. In essence, if an image is presented as input, the number of neurons at the first layer will be the same as the number of pixels of the input image. Therefore, if an image was used as input of a fully-connected network, it would require a combinatorial number of connections between neurons and hence the training of this network would be unmanageable. CNNs are capable of dealing with the computational complexity issues by connecting sub-region of previous layers to neurons and the weights and bias are shared allowing to look for the same feature in several regions. In the second layer, each neuron is connected to a subset of neurons from the previous layer, called receptive field. This way, receptive fields of neurons in deeper layers involve a combination of receptive fields from several neurons from the previous layers.

ImageNet Data Set

ImageNet is a large public visual data set of over 15 million labeled images taking part of about 22 thousand categories. The annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) started in 2010 and uses a subset of ImageNet formed by roughly 1000 images in each of the 1000 categories[2]. The ILSVRC 2012 data set [179] was previously divided into training, validation and test images. The validation and test data consist of 50000 and 100000 annotated images but only validation labels data were released. The remaining images (test data) were released without label and will be used to evaluate the algorithm. Since this data set was part of a competition, the participants had to submit their results on the available test images and only at the end of the competition they knew the results and the respective winner. These 150000 images (validation and test) were not part of the training data that is formed by 1.2 million images containing the 1000 categories. The challenge consisted of three tasks and the data set [179] was already divided and publicly available.

Pre-trained Models

Train a network from scratch using a large amount of color images is computationally expensive and time consuming. Thereby, there are some pre-trained Convolutional Network (ConvNet) models available at Caffe [106] Model Zoo. In this section, an explanation is given on the different architectures of several pre-trained models and some preliminary results available on Model Zoo are shown[3].

CaffeNet/AlexNet Krizhevsky's work [117] presents a DCNN constituted by five convolutional and three fully-connected layers called AlexNet model. The convolutional layers are followed by a ReLU layer, then the neurons are normalized by a Local Response Normalization (LRN) layer and finally a down-sampling is performed by a max-pooling layer. The fully-connected layers are followed by a ReLU and a Dropout layer with dropout ratio of 0.5.

Two techniques were proposed to deal with overfitting: first, to artificially increase the data set by applying small transformations to the original images like translations and horizontal reflections or change intensity of color channels during training and secondly, use the dropout technique.

[2]source: http://image-net.org/challenges/LSVRC/2012/browse-synsets [seen in November, 2016]
[3]source: http://caffe.berkeleyvision.org/model_zoo.html [seen in November, 2016]

Caffe [106] provides a reference CaffeNet[4]model which is a modification of AlexNet where the order of Pooling and Normalization (LRN) layers are switched. Besides this, all the rest remains the same including all the parameters of all layers. The change originates a slight computational advantage to CaffeNet since the max-pooling operation is done before the normalization which will use less memory and calculations. Yet, there is not a significant performance difference between both models.

A pre-trained version of both models is available and both were tested to check for performance differences (see Table A.1). Both models were trained without the data-augmentation used to prevent the overfit mentioned on [117] and the AlexNet model was initialized with non-zero biases of 0.1 instead of 1.[5] Results released at [117] show a top-1 classification error of 40.7% and a top-5 classification error of 18.2% of AlexNet model while public replication of AlexNet presented a top-1/top-5 classification error of 42.9% / 19.8%. The results of CaffeNet differed by less than 0.5% from the AlexNet but once it requires less memory, the CaffeNet was the chosen model to perform the tests.

GoogLeNet GoogLeNet is a deep convolutional neural network with 22 weight layers proposed by Szegedy *et al.* [197] for classification and detection tasks which improved the use of computational resources. It has nine Inception modules that allow parallel pooling and convolution operations. For classification, it uses the spatial average of the feature maps from the last convolution layer as the confidence of categories via a global average pooling layer. The resulting vector is then used as input into the softmax layer. The most direct form of improving the performance of deep networks is by increasing their size including depth (more layers) and width (more units at each layer). Even with a bigger network, a constant computational budget was managed by using additional 1×1 convolutions as dimension reduction method [134] before the expensive 3×3 and 5×5 convolutions and by replacing fully connected layers by sparse ones. A replication of the model in [197] was trained and the weights file is publicly available[6]. However, there are some training differences that should be highlighted: the replication uses "xavier" to initialize the weights instead of "gaussian"; the learning rate decay policy is different allowing a faster training and training was done without data-augmentation. Xavier initialization is characterized by setting the weights with a Gaussian distribution with zero mean and a weight variance equal to the inverse of the number of input neurons ensuring faster convergence [71].
On one hand, the original model [197] achieved a top-5 classification error of 10.07% in the validation data and a localization error of 38.02%. The top-1 classification error was not disclosed. On the other hand, replication model obtained a top-1 error of 31.3% and a top-5 error of 11.1%. The localization error was not published. Once the weights file of the replication model was the one used, the results obtained on this project were compared with theirs (see Table A.1).

VGGNet It is a DCNN for object recognition developed and trained by Oxford's renowned Visual Geometry Group (VGG)[7] [189]. This architecture was developed with the purpose of exploring the effect of the ConvNet depth on its accuracy. Different configurations were used that goes from a ConvNet with 11 weight layers to a ConvNet with 19 weight layers and the performance of individual ConvNet models were evaluated. For localization task, the 16 weight layers architecture was used where the last fully connected layer predicts the bounding box location instead of the class scores.

In comparison with the state-of-the-art at the time, an evident improvement was reached with a deeper network achieving the optimal configuration at 16-19 weight layers. Since usually deeper networks mean more parameters and more chance to overfit, Simonyan *et al.* used small 3×3 filters in all convolutional layers. Besides this improvement, a demonstration of the generalization power of their model was done by achieving the state-of-the-art results with other image recognition data sets such as PASCAL Visual Object Classes (2007 and 2012) [51]. The

[4]source: https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet [seen in December, 2016]
[5]source: https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet [seen in December, 2016]
[6]source: https://github.com/BVLC/caffe/tree/master/models/bvlc_googlenet [seen in December, 2016]
[7]source: https://github.com/BVLC/caffe/wiki/Model-Zoo#models-used-by-the-vgg-team-in-ilsvrc-2014 [seen in December, 2016]

16 weight layer configuration achieved a top-1/top-5 classification error of 25.6% / 8.1% and a localization error of 26.9%. The 19 weight layer configuration decreased only 1% of the previous classification error which proved to be the best results achieved so far. In this project, the pre-trained model VGGNet that was used has 16 weight layers.

Table A.1 has a compilation of the classification and localization errors disclosed by the current state-of-the-art. There are some fields of the table that contain a line which means that these results have not been published. As explained on Section A, AlexNet pre-trained model is not used in our tests once there is no significant difference of performance between AlexNet and CaffeNet pre-trained model and CaffeNet requires less memory.

Table A.1: ConvNet performance following the state of the art.

Model	Number of weight layers	**Classification Error** Top-1 [%]	Top-5 [%]	**Localization Error** Error [%]
CaffeNet [117]	8	42.6	19.6	—-
AlexNet [117]	8	42.9	19.8	—-
GoogLeNet [197]	22	31.3	11.1	38.02
GoogLeNet Feedback [34]	–	30.5	10.5	38.80
VGGNet [188]	8	39.7	17.7	44.60
VGGNet [189] (16 layers)	16	25.6	8.1	26.90
VGGNet [189] (19 layers)	19	25.5	8.0	—-